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, 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

ACM Reference format:
Julia Kiseleva and Maarten de Rijke. 2017. Evaluating Personal Assistants on Mobile devices. In *Proceedings of The First International Workshop on Conversational Approaches to Information Retrieval, Tokyo, Japan, August 11, 2017 (CAIR 2017),* 4 pages.

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,[1] and it overtook the usage of desktops in October 2016. To a large degree, this increase is due to the 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 are the recent significant improvements 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 due to gesture- and voice-based control, such as 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. This 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 a personal assistant? Previously, a common practice for evaluation was to create a 'gold' standard[2] [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, 31, 66] or a user's

[1] http://gs.statcounter.com/#desktop+mobile+tablet-comparison-ww-monthly-201208-201612
[2] A gold standard is a set of "correct" answers as judged by editorial judges.

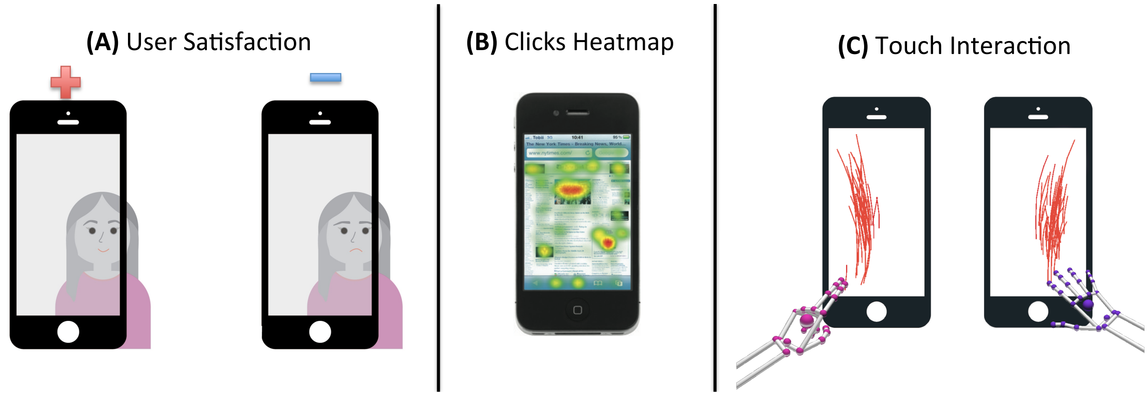

Figure 1: Illustrations of (A) how to define user satisfaction, (B) where users click on the mobile screen, and (C) how touches are tracked

past preferences [25, 35, 68]. User satisfaction is widely adopted as a subjective measure of the quality of the search experience [30]. It has become common practice to evaluate personal assistants on desktops by analysing interaction signals such as clicks (if users like a result, 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 from 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 mobile devices 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, 31, 32, 36, 37, 53, 67];
- sensitive to changes in the outside world [33, 34].

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 detail, allowing us to reason about how a user's current environment impacts his satisfaction with personal assistants.

How can we 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 be 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 is analyzing gestures the way to infer context-aware user satisfaction? Analyzing click heat maps as displayed in Figure 1(B), is quite tricky. Because the screen size is small, it is difficult to click an item, and conversely, not to click inadvertently. Analyzing gesture-based patterns is a better way to infer user satisfaction as it helps to decipher hidden behavioral aspects, e.g., swipes in the two figures show in Figure 1(C) clearly belong to left- and right-handed people. 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 interactions 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.

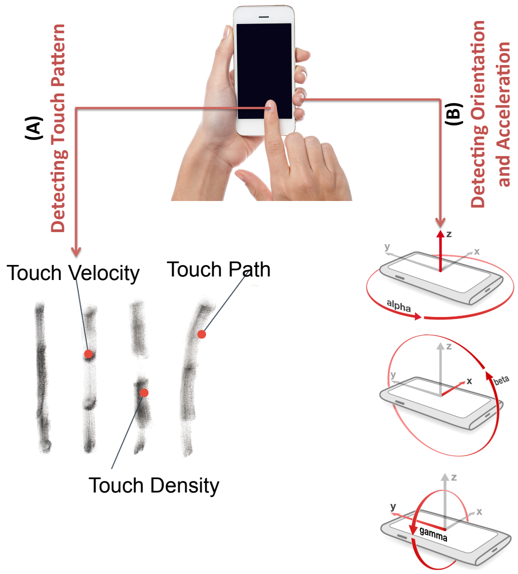

Figure 2: Modelling gesture-based interactions.

3 DISCUSSION

We propose ventures into a new area of research, by moving beyond the interaction models that have been shown to be effective for desktop applications and fully embrace the new paradigm of gesture- and voice-controlled personal assistants on mobile devices. Such a move has the potential to impact billions of users around the world, and it has a large scientific impact. The research we propose will significantly contribute to our scientific understanding of user satisfaction in mobile environments, and to the value of massive-scale gesture-based interaction logs to infer user satisfaction based on complex and subtle interaction patterns. Obtaining insights into this value is crucial if we want to evaluate new algorithms for search and recommendation in a mobile or screen-less environment.

RQ1: *How to model interaction with gesture- and voice-controlled devices?* We need to encompass all gesture- and voice-based features related users' interactions 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].[3] 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., 'shakes.' 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 [31, 32, 37] and changing environments [33, 34]. In addition to unsupervised logs, dedicated user experiments should be conducted to gather rich, explicitly annotated data for further analysis and validation.

[3]The viewport is the visible region on the device.

RQ3: *How to predict context-aware user satisfaction with personal assistants using gesture-based signals on mobile devices?* Recently, deep neural networks have given rise to significant performance improvements in speech recognition [24] and computer vision tasks [45]. They have also led to exciting breakthroughs in novel application areas such as automatic voice translation [46], image captioning [62], and conversational assistants [19, 61]. However, there are only few publications on using deep neural networks to model user interaction behavior. So far these have been confined to desktop settings [9, 10, 49], where the advantage of neural approaches has been clearly demonstrated; the time is right to put them to work in a mobile setting.

Acknowledgments. This research was supported by Ahold Delhaize, Amsterdam Data Science, the Bloomberg Research Grant program, the Criteo Faculty Research Award program, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Microsoft Research Ph.D. program, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.001.-116, HOR-11-10, CI-14-25, 652.002.001, 612.001.551, 652.001.003, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. *CARS*, 2010.
[2] E. Agichtein, E. Brill, and S. T. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, 2006.
[3] A. O. Alves and F. C.Pereira. Making sense of location context. In *ContextDD*, 2012.
[4] M. Bachynskyi, A. Oulasvirta, G. Palmas, and T. Weinkauf. Is motion capture-based biomechanical simulation valid for hci studies?: study and implications. In *CHI*, pages 3215–3224, 2014.
[5] E. Bakshy and D. Eckles. Uncertainty in online experiments with dependent data: an evaluation of bootstrap methods. In *KDD*, pages 1303–1311, 2013.
[6] J. Bergstrom-Lehtovirtaa and A. Oulasvirta. Modeling the functional area of the thumb on mobile touchscreen surfaces. In *SIGCHI*, pages 1991–2000, 2014.
[7] D. Bernhardt. *Emotion inference from human body motion.* PhD thesis, University of Cambridge, UK, 2010.
[8] D. Bernhardt and P. Robinson. Detecting emotions from connected action sequences. In *IVIC*, pages 1–11, 2009. doi: 10.1007/978-3-642-05036-7_1.
[9] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A neural click model for web search. In *WWW*, pages 531–541. International World Wide Web Conferences Steering Committee, 2016.
[10] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A context-aware time model for web search. In *SIGIR*, pages 205–214, 2016.
[11] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, 2010.
[12] A. Chuklin, P. Serdyukov, and M. de Rijke. Click model-based information retrieval metrics. In *SIGIR*, pages 493–502, 2013.
[13] A. Chuklin, I. Markov, and M. de Rijke. *Click Models for Web Search.* Morgan & Claypool Publishers, 2015. ISBN 1627056475, 9781627056472.
[14] A. Deng, T. Li, and Y. Guo. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *WWW*, pages 609–618, 2014.
[15] A. Deng, J. Lu, and J. Litz. Trustworthy analysis of online a/b tests: Pitfalls, challenges and solutions. In *WSDM*, 2017.
[16] A. Drutsa, G. Gusev, and P. Serdyukov. Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *WSDM*, pages 27–36, 2015.

[17] A. Drutsa, G. Gusev, and P. Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW*, pages 256–266, 2015.

[18] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais, and T. White. Evaluating implicit measures to improve web search. *TOIS*, 23(2):147–168, 2005.

[19] D. Griol and Z. Callejas. A neural network approach to intention modeling for user-adapted conversational agents. *Comp. Int. and Neurosc.*, 2016:8402127:1–8402127:11, 2016.

[20] Q. Guo and E. Agichtein. Ready to buy or just browsing? detecting web searcher goals from interaction data. In *SIGIR*, pages 130–137, 2010.

[21] Q. Guo and E. Agichtein. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW*, pages 569–578, New York, New York, USA, 2012. ACM Press.

[22] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In *CIKM*, pages 2050–2054, 2012.

[23] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: User behavior as a predictor of a successful search. In *WSDM*, pages 221–230, 2010.

[24] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, 2012.

[25] Y.-C. Ho, Y.-T. Chiang, and J. Hsu Yung-Jen. Who likes it more?: mining worth-recommending items from long tails by modeling relative preference. In *WSDM*, pages 253–262, 2014.

[26] K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10(1):1–117, 2016.

[27] J. Huang and A. Diriye. Web user interaction mining from touch enabled mobile devices. In *HCIR Workshop*, 2012.

[28] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

[29] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, 2005.

[30] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *FTIR*, 3(1-2):1–224, 2009.

[31] J. Kiseleva, H. T. Lam, M. Pechenizkiy, and T. Calders. Predicting current user intent with contextual markov models. In *ICDMW*, pages 391–398, 2013.

[32] J. Kiseleva, H. T. Lam, M. Pechenizkiy, and T. Calders. Discovering temporal hidden contexts in web sessions for user trail prediction. In *WWW Workshop TempWeb*, pages 1067–1074, 2013.

[33] J. Kiseleva, E. Crestan, R. Brigo, and R. Dittel. Modelling and detecting changes in user satisfaction. In *CIKM*, pages 1449–1458, 2014.

[34] J. Kiseleva, J. Kamps, V. Nikulin, and N. Makarov. Behavioral dynamics from the SERP's perspective: What are failed SERPs and how to fix them? In *CIKM*, pages 1561–1570, 2015.

[35] J. Kiseleva, M. J. I. MÃijller, L. Bernardi, C. Davis, I. Kovacek, M. Stafseng Einarsen, J. Kamps, A. Tuzhilin, and D. Hiemstra. Where to go on your next trip? optimizing travel destinations based on user preferences. In *SIGIR*, pages 1097–1100, 2015.

[36] J. Kiseleva, A. Montes García, J. Kamps, and N. Spirin. The impact of technical domain expertise on search behavior and task outcome. In *Proceedings of WSDM Workshop QRUMS*, 2016.

[37] J. Kiseleva, A. Tuzhilin, J. Kamps, M. J. I. Müller, L. Bernardi, C. Davis, I. Kovacek, M. Stafseng Einarsen, and D. Hiemstra. Beyond movie recommendations: Solving the continuous cold start problem in e-commercerecommendations. *CoRR*, 1607.07904, 2016.

[38] J. Kiseleva, K. Williams, A. H. Awadallah, I. Zitouni, A. Crook, and T. Anastasakos. Predicting user satisfaction with intelligent assistants. In *SIGIR*, 2016.

[39] J. Kiseleva, K. Williams, J. Jiang, A. H. Awadallah, I. Zitouni, A. Crook, and T. Anastasakos. Understanding user satisfaction with intelligent assistants. In *CHIIR*, pages 121 – 130, 2016.

[40] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *KDD*, pages 1857–1866, 2014.

[41] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[42] D. Lagun and E. Agichtein. Inferring searcher attention by jointly modeling user interactions and content salience. In *SIGIR*, pages 483–492, 2015.

[43] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR*, pages 113–122, 2014.

[44] D. Lagun, D. McMahon, and V. Navalpakkam. Understanding mobile searcher attention with rich ad formats. In *CIKM*, pages 599–608, 2016.

[45] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[46] W. D. Lewis. Skype translator: Breaking down language and hearing barriers. *Proceedings of Translating and the Computer (TC37)*, 2015.

[47] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR*, pages 493–502, 2015.

[48] M. F. McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169, 2002.

[49] B. Mitra. Exploring session context using distributed representations of queries and reformulations. In *SIGIR*, pages 3–12. ACM, 2015.

[50] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna. Quality estimation for automatic speech recognition. In *COLING*, pages 1813–1823, 2014.

[51] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci. In *SIGCHI*, pages 919–928, 2005.

[52] Z. Ren, J. Jia, Q. Guo, K. Zhang, and L. Cai. Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data. In *ICME*, 2014.

[53] S. Rendle, Z. Gantner, C. Freudenthaler, and L. Schmidt-Thieme. Fast context-aware recommendations with factorization machines. In *SIGIR*, pages 635–644, 2011.

[54] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI*, pages 2997–3002, 2008.

[55] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[56] M. Shokouhi, U. Ozertem, and N. Craswell. Did you say u2 or youtube? inferring implicit transcripts from voice search logs. In *WWW*, pages 1215–1224, 2016.

[57] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD*, pages 17–26, Washington, DC, 2010.

[58] G. Tür. Extending boosting for large scale spoken language understanding. *Machine Learning (ML)*, 69(1):55–74, 2007.

[59] G. Tür, Y.-Y. Wang, and D. Z. Hakkani-Tür. Techware: Spoken language understanding resources [best of the web]. *IEEE Signal Process. Mag. (SPM)*, 30(3): 187–189, 2013.

[60] G. Tür, Y.-Y. Wang, and D. Z. Hakkani-Tür. Understanding spoken language. *Computing Handbook*, 3rd ed(41):1–17, 2014.

[61] O. Vinyals and Q. V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015. URL http://arxiv.org/abs/1506.05869.

[62] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[63] M. Weigel, T. Lu, G. Bailly, A. Oulasvirta, C. Majidi, and J. Steimle. Iskin: flexible, stretchable and visually customizable on-body touch sensors for mobile computing. In *SIGCHI*, pages 2991–3000, 2015.

[64] K. Williams, J. Kiseleva, A. Crook, I. Zitouni, A. H. Awadallah, and M. Khabsa. Is this your final answer? evaluating the effect of answers on good abandonment in mobile search. In *SIGIR*, 2016.

[65] K. Williams, J. Kiseleva, A. C. Crook, I. Zitouni, A. H. Awadallah, and M. Khabsa. Detecting good abandonment in mobile search. In *WWW*, pages 495 – 505, 2016.

[66] L. Wu, A. Chin, Y. Zhou, X. Wang, Y. Guo, and J. Li. Context-aware prediction of user's first click. In *ContextDD*, 2012.

[67] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, and H. Li. Context-aware ranking in web search. In *SIGIR*, 2010.

[68] L. Xiang, Q. Yuan, S. Zhao, L. Chen, X. Zhang, Q. Yang, and J. Sun. Temporal recommendation on graphs via long- and short-term preference fusion. In *KDD*, pages 723–732, 2010.

[69] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *CIKM*, pages 1561–1564, 2010.

[70] K. S. Young, C. E. Parsons, R. T. LeBeau, B. A. Tabak, A. R. Sewart, A. Stein, M. L. Kringelbach, and M. G. Craske. Sensing emotion in voices: Negativity bias and gender differences in a validation study of the oxford vocal (âĂŸoxvocâĂŹ) sounds database. *American Psychological Association*, 2016.

[71] H. Zhu, E. Chen, H. Xiong, K. Yu, H. Cao, and J. Tian. Mining mobile user preferences for personalized context-aware recommendation. *TIST*, 5(4):58, 2015.