Designing Multimodal Self-Tracking Technologies to Promote Data Capture and Self-Reflection

Yuhan Luo
yuhanluo@umd.edu
University of Maryland
College Park, MD, USA

ABSTRACT

Self-tracking is a powerful means to help individuals monitor and improve their behaviors. While numerous tracking technologies are available, it has been challenging to lower the tracking burden whilst promoting reflection. This is because low-burden tracking technologies utilizing automated sensors reduce people's awareness of their data; reflective tracking approaches, such as manual typing, often impose a high data capture burden. Motivated by speech input's fast and expressive nature, my dissertation examines how speech input complements traditional touch input in supporting self-tracking. Taking a research-through-design approach, I examine the use of speech input in exercise tracking and food journaling, and evaluate a flexible multimodal self-tracking system incorporating touch and speech input in a domain-agnostic context. I hope this work can inform the design of multimodal technologies to support low-burden, engaging, and reflective self-tracking experiences.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **User centered design**; **Sound-based input / output**; **Field studies**; **Participatory design**; **Ubiquitous and mobile computing design and evaluation methods**.

KEYWORDS

Self-tracking, personal informatics, speech input, multimodal interaction, exercise, food journaling

ACM Reference Format:
Yuhan Luo. 2021. Designing Multimodal Self-Tracking Technologies to Promote Data Capture and Self-Reflection. In *Designing Interactive Systems Conference 2021 (DIS '21 Companion), June 28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3468002.3468232

1 CONTEXT AND MOTIVATION

Self-tracking, a practice of capturing one's personal data such as exercise and food, is powerful to enhance self-awareness, enable reflection, and yield positive behavior change [7]. With the advances in sensing technologies (e.g., Fitbit [14]), many types of personal data can be captured automatically (e.g., steps, sleep, heart rate), but such automated tracking reduces individuals' engagement with data collection, limiting the level of self-awareness and reflection [13]. Manual tracking (e.g., typing, handwriting), on the other hand, can collect rich contexts and increase situated awareness, but imposes high data capture burden [3, 6].

In recent years, speech input has been rapidly integrated in our daily life—it is embedded in various interfaces, agents, and environments [41]. In particular, I see the potential of speech input for facilitating self-tracking in four aspects. First, people speak faster than they type through touch or keyboard input [37]. By capturing the same amount of information faster, speech input can lower the data capture burden. Second, people tend to be expressive when they talk [5], so that speech input can help collect rich details that might otherwise be overlooked through traditional manual input. Third, speaking through natural languages enables people to describe their activities and thoughts in a flexible manner (e.g., time-related queries [23]). Lastly, the interaction with some speech-enabled applications is hands-free (e.g., Google Assistant[17], Siri [2]), allowing people to multitask while not having to concentrate on their device screen. Such hands-free interaction also makes the system accessible to broader populations [34].

Despite its potential, speech input is limited in supporting people editing their data on the fly, and may raise privacy concerns [30]. Current research provide little understanding on whether and how speech input can be useful in self-tracking contexts, including how it supports capturing structured versus unstructured personal data, how it works with other input modalities, and when speech-enabled data capture is favorable in what contexts.

In this light, my dissertation first explores the use of speech input in two specific self-tracking contexts: exercise and food, which are essential aspects of human health [33]. I then examine how people adopt speech and touch input in a domain-agnostic context where they can customize their own trackers. With the lessons learned, my overarching goal is to inform the design of low-burden, engaging, and reflective self-tracking technologies combing speech with other input modalities.

2 RESEARCH QUESTIONS

Specifically, I aim to answer four research questions:
RQ1. How does a smart speaker complement a mobile app in supporting exercise training and tracking?
RQ2. To facilitate working with dietitians, how does multimodal input support customizing food trackers for patients with different dietary problems?
RQ3. How does speech input support capturing daily food practice regarding data richness and data capture burden?
RQ4. Given the flexibility to customize one's own trackers, how do

DIS '21 Companion, June 28-July 2, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8559-6/21/06.
https://doi.org/10.1145/3468002.3468232

people adopt touch and speech input to capture their personal data?

RQ1 focuses on examining the roles of smart speakers and mobile apps in supporting exercise tracking. RQ2 and RQ3 aim to identify design opportunities for multimodal food journaling. RQ4 goes beyond a predefined domain to examine how people adopt speech and touch input to capture their personal data.

3 RELATED WORK

People naturally interact with the world through multiple communication channels, among which vision and speech are most common [4, 35]. While screen-based touch input has been the mainstream for Human-Computer Interaction (HCI), we have seen a growing uptake of speech input in recent years [19, 39]. Below, I describe related on general approaches for speech-based data collection, together with commercial tools and research on exercise tracking and food journaling that involved multimodal interaction. A more comprehensive review can be found in [27–29].

3.1 Speech-Based Data Collection

With the rise of speech recognition and natural language processing (NLP), researchers have explored the opportunities for speech-based data collection [9, 11, 36, 40]. For example, while comparing speech with text input in responding to survey questions, Revilla and colleagues found that participants who used speech input spent less time and provided more elaborated answers than those who used text [36], but Schober and colleagues found that speech input generated less precise responses regarding numbers [40].

Another field that has explored speech-based data capture is clinical data entry, where doctors have to enter patients' data such as medical reports [1, 12, 32, 38, 43, 44]. Wenzel and colleagues showed that compared with handwriting and keyboard typing, doctors rated speech dictation with a higher level of satisfaction [44]. Although prior work suggested speech input's benefits and limitations in capturing objective research data, there is a lack of empirical understanding on how speech input supports capturing personally-attached and context-dependent self-tracking data.

3.2 Health Applications on Smart Speakers

As smart speakers become a part of our interaction routines, various speech-enabled applications on smart speakers have emerged to support daily activities, including health and fitness. In 2017, researchers found 309 speech-enabled applications under the "Health & Fitness" category among Amazon Alexa skills and Google Actions [8], which has reached over 2,200 as of May 2021. Many of these applications support fitness training (e.g., 7-Minute Workout [18]) by guiding people through a set of workouts with background music, or allow people to ask questions about their data captured by other devices (e.g., Fitbit skill [15]), but do not support capturing exercise data using speech.

3.3 Multimodal Food Journaling

Food journaling is known to be burdensome due to the complexity of meal composition and variation in preparation methods [10]. Therefore, researchers have explored different input modalities, including photos [26], barcode [21], accelerated search [20], and smart sensors [31], to lower the data capture burden. Korpusik and colleagues incorporated speech into food journaling by developing Coco Nutrition [24], a conversation-based calorie counter that captures natural languages through speech or text input, and automatically calculates the calorie consumption [25]. Similarly, a commercial app Talk-to-Track [16] takes people's speech input and converts it into calorie information. While these approaches largely focus on capturing food nutrients leveraging speech input's convenience, they often ignore the broader eating contexts (e.g., mood, eating environment) beyond what people eat, which are valuable information to increase situated awareness [3, 45].

Besides exercise tracking and food journaling, I believe that speech input can enable fast, expressive, and flexible data collection in many other self-tracking contexts. For example, giving people the flexibility to decide what data to capture with different input format, researchers found that they often use speech input to capture contextual information such as reasons of coffee intake [22].

4 RESEARCH APPROACHES

4.1 TandemTrack: Exercise Tracking on Smart Speaker & Mobile Phone

To answer RQ1, I designed and developed TandemTrack, a multimodal system comprised of an Alexa skill and a mobile app (Figure 1), which supports a simple exercise regimens (alternating between sit-ups and push-ups), data capture, feedback, and daily reminders. I deployed TandemTrack to 22 participants in a four-week between-subjects study, with one group using the mobile app only and the other group using both the app and the skill. Although the two groups did not differ in their exercise adherence and performance, the findings highlighted the benefits of hands-free interaction in supporting exercise training and data capture, especially for push-ups that requires intensive use of hands. When it came to choosing between the app and the skill, participants' decisions were influenced by various environmental and social factors, including their personal preferences, proximity to Amazon Echo, and people around them. A research paper describing the study results has been published at CHI 2020 [28].

4.2 Co-Designing with Dietitians

To answer RQ2, I conducted individual co-design workshops with six registered dietitians. During the workshops, each participant was asked to describe two representative patient persona, and to create food trackers for those personas using paper-based design widgets. The 12 patient personas had common attributes while demonstrating unique characteristics in terms of age, dietary problems, symptoms, and goals. As a result, dietitians generated diverse and individualized tracker designs regarding what to track, when to track, and how to track (See Figure 2 for examples). Particularly of interest, dietitians suggested that speech input could encourage people to record more frank thoughts with its ephemeral nature, thereby promoting situated reflection on one's eating behaviors. This finding inspired my next study on understanding the experience of speech-enabled food journaling. A research paper describing the study results has been published at CHI 2019 [29].

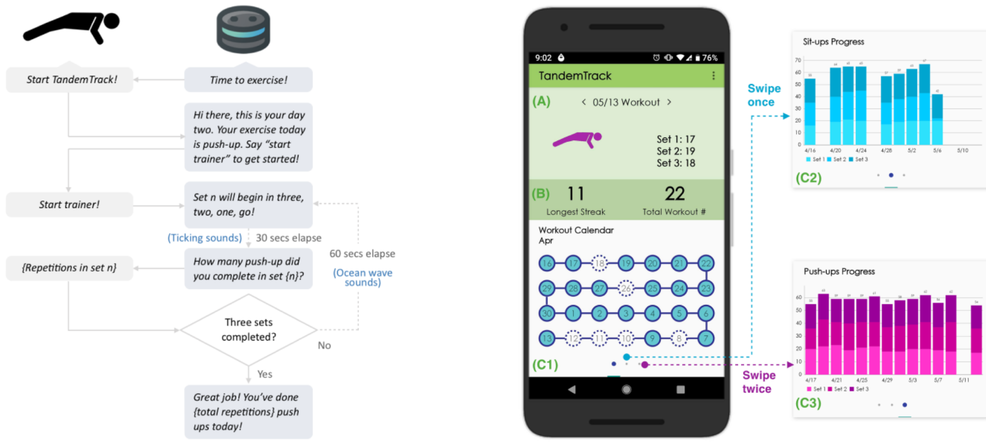

Figure 1: An example of using the TandemTrack Alexa skill to do push-ups (left); the home screen of the TandemTrack mobile app (right): the daily exercise feedback (A); a summary of the longest streak and complete exercise sessions (B); a series of aggregated feedback—exercise streak view (C1), sit-up progress (C2), and push-up progress (C3).

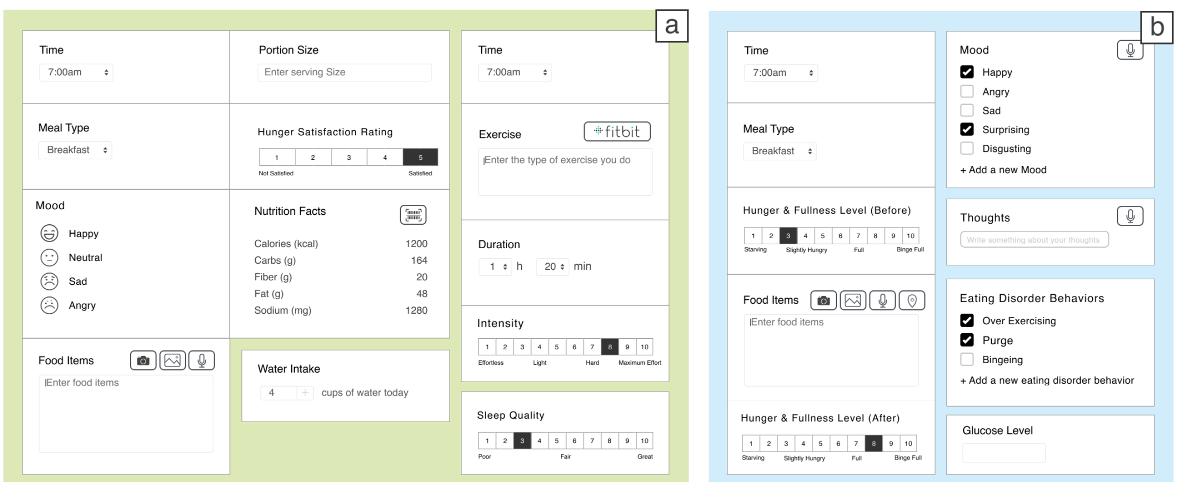

Figure 2: The digitized version of paper-based food trackers for two different patient personas: a weight management patient (a) and an eating disorder patient with diabetes (b). Items grouped together are meant to be tracked together at the same time. Icons next to the title represent alternative ways to capture the information (e.g., taking a photo is an alternative way to capture food items).

4.3 FoodScrap: Speech-Based Food Journaling

To answer RQ3, I created FoodScrap, a mobile app that takes audio recording as responses to a set of guided questions on people's daily food practice, including food components, preparation methods, and food decisions (Figure 3). I conducted a one-week data collection study by deploying FoodScrap to 11 participants from diverse food cultures, followed by a post-study questionnaire on User Burden Scale [42] and debriefing interviews. With speech input, participants not only detailed the ingredients in their meals and steps of preparation procedures, but also elaborated their food decisions beyond what we asked in FoodScrap. More importantly, participants recognized speech input as a reflection tool that encouraged them to think aloud. A research paper describing the study results has been published at DIS 2021 [27].

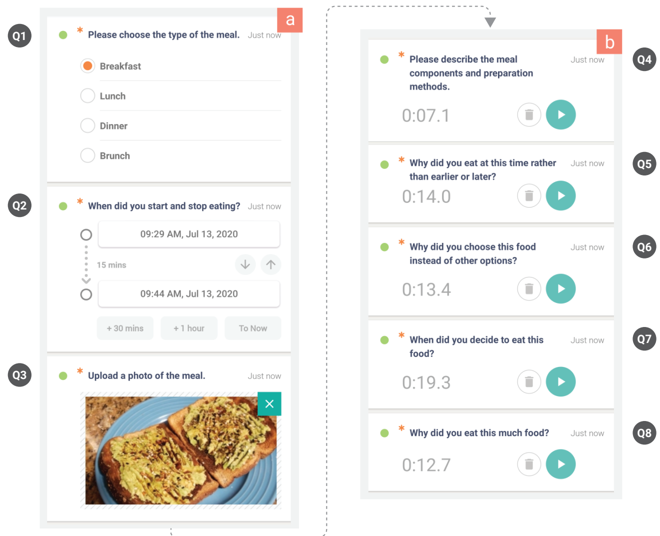

Figure 3: The data capture screen of FoodScrap: (a) questions on meal type, eating duration, and photo of the meal; (b) questions on meal components, preparation methods, and food decisions.

5 NEXT STEP

While my prior research showed how speech input helped to lower data capture burden while collecting rich details, it remains unclear whether these findings can be generalized to other self-tracking contexts, and how we can tailor the input modalities to meet people's various information needs [29].

To answer RQ4, I plan to build a multimodal self-tracking system by extending OmniTrack [22]. Besides supporting people customizing their own trackers, the updated system will enable a "global speech input" entry, allowing people to capture multiple data fields using natural languages (e.g., log the time and activity information from utterance "*I had a cup of coffee three hours ago*"). In addition, people can edit each data field individually using touch or speech input. I plan to deploy the system with 20 participants in a two-week long study, focusing on how people describe their personal data in natural languages, and how they choose between speech and touch input to capture different types of personal data.

6 EXPECTED CONTRIBUTIONS

The contributions of my dissertation to the Human-Computer Interaction (HCI) community are three folds, including:
(1) The design and implementation of a multimodal self-tracking system that allows people to capture different types of personal data using speech or touch input.
(2) An empirical understanding of the benefits and challenges of adopting speech input in various self-tracking contexts.
(3) Design implications for incorporating speech to build multimodal self-tracking systems that lower the data capture burden, enrich tracking experience, and foster reflection.

REFERENCES

[1] Mohd Khanapi Abd Ghani and Ika Novita Dewi. 2012. Comparing speech recognition and text writing in recording patient health records. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*. IEEE, 365–370. https://doi.org/10.1109/IECBES.2012.6498100
[2] Apple, Inc. 2021. *Siri*. Retrieved Accessed: 2021-05-07 from https://www.apple.com/siri/
[3] Johnna Blair, Yuhan Luo, Ning F Ma, Sooyeon Lee, and Eun Kyoung Choe. 2018. OneNote Meal: A photo-based diary study for reflective meal tracking. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 252. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371351/
[4] Harry Bunt, Robbert-Jan Beun, and Tijn Borghuis. 1998. *Multimodal human-computer communication: systems, techniques, and experiments*. Vol. 1374. Springer Science & Business Media.
[5] Barbara L Chalfonte, Robert S Fish, and Robert E Kraut. 1991. Expressive richness: a comparison of speech and text as media for revision. In *Proceedings of the 1991 CHI Conference on Human Factors in Computing Systems*. 21–26. https://doi.org/10.1145/108844.108848
[6] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 121–132. https://doi.org/10.1145/2750858.2804266
[7] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*. 1143–1152. https://doi.org/10.1145/2556288.2557372
[8] Arlene E Chung, Ashley C Griffin, Dasha Selezneva, and David Gotz. 2018. Health and Fitness Apps for Hands-Free Voice-Activated Assistants: Content Analysis. *JMIR mHealth and uHealth* 6, 9 (2018), e174. http://doi.org/10.2196/mhealth.9705
[9] Frederick G Conrad, Michael F Schober, Christopher Antoun, H Yanna Yan, Andrew L Hupp, Michael Johnston, Patrick Ehlen, Lucas Vickers, and Chan Zhang. 2017. Respondent mode choice in a smartphone survey. *Public Opinion Quarterly* 81, S1 (2017), 307–337. https://doi.org/10.1093/poq/nfw097
[10] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty. 2015. Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. 1159–1162. https://doi.org/10.1145/2702123.2702155
[11] Marika De Bruijne and Arnaud Wijnant. 2013. Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review* 31, 4 (2013), 482–504. https://doi.org/10.1177/0894439313483976
[12] Yaron D Derman, Tamara Arenovich, and John Strauss. 2010. Speech recognition software and electronic psychiatric progress notes: physicians' ratings and preferences. *BMC medical informatics and decision making* 10, 1 (2010), 1–7. https://doi.org/10.1145/1497185.1497286

[13] Xianghua Ding, Shuhan Wei, Xinning Gui, Ning Gu, and Peng Zhang. 2021. Data Engagement Reconsidered: A Study of Automatic Stress Tracking Technology in Use. In *Proceedings of the 2021 CHI conference on human factors in computing systems*.

[14] Fitbit. 2021. Fitbit. Retrieved Accessed: 2021-05-07 from https://www.fitbit.com/

[15] Fitbit. 2021. Fitbit Skill. Retrieved Accessed: 2021-05-07 from https://www.amazon.com/dp/B01CH4BP28

[16] Genesant Technologies, Inc. 2021. Talk-to-Track. Retrieved Accessed: 2021-05-07 from https://www.talktotrack.com/

[17] Google, Inc. 2021. *Google Assistant.* Retrieved Accessed: 2021-05-07 from https://assistant.google.com/

[18] Health Xperts Production. 2021. Seven-Min Fitness Challenge. Retrieved Accessed: 2021-05-07 from http://7minworkoutapp.net/

[19] Xiaochen Huang and Emmanuel Agu. 2015. A speech-based mobile app for restaurant order recognition and low-burden diet tracking. In *ICSH*. Springer, 333–339. https://doi.org/10.1007/978-3-319-29175-8_31

[20] Jisu Jung, Lyndal Wellard-Cole, Colin Cai, Irena Koprinska, Kalina Yacef, Margaret Allman-Farinelli, and Judy Kay. 2020. Foundations for systematic evaluation and benchmarking of a mobile food logger in a large-scale nutrition study. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–25. https://doi.org/10.1145/3397327

[21] Yeong Gug Kim and Eunju Woo. 2016. Consumer acceptance of a quick response (QR) code for the food traceability system: Application of an extended technology acceptance model (TAM). *Food Research International* 85 (2016), 266–272. https://doi.org/10.1016/j.foodres.2016.05.002

[22] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A flexible self-tracking approach leveraging semi-automated tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–28. https://doi.org/10.1145/3130930

[23] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@ Hand: Fostering Visual Exploration of Personal Data on Smartphones Leveraging Speech and Touch Interaction. *arXiv preprint arXiv:2101.06283* (2021).

[24] Mandy Korpusik and Jim Glass. 2021. Coco Nutrition. https://www.coco-nutrition.com/.

[25] Mandy Korpusik, Salima Taylor, Sai Krupa Das, Cheryl Gilhooly, Susan Roberts, and James Glass. 2019. A Food Logging System for iOS with Natural Spoken Language Meal Descriptions (P21-009-19). *Current developments in nutrition* 3, Supplement_1 (2019), nzz041–P21. https://doi.org/10.1093/cdn/nzz041.P21-009-19

[26] Brian Y Lim, Xinni Chng, and Shengdong Zhao. 2017. Trade-off between automation and accuracy in mobile photo recognition food logging. In *Proceedings of the Fifth International Symposium of Chinese CHI*. 53–59. https://doi.org/10.1145/3080631.3080640

[27] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. FoodScrap: Promoting Rich Data Capture and Reflective Food Journaling Through Speech Input.. In *Proceedings of the 2021 Conference on Designing Interactive System*. ACM. https://doi.org/10.1145/3461778.3462074

[28] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3313831.3376616

[29] Yuhan Luo, Peiyi Liu, and Eun Kyoung Choe. 2019. Co-Designing food trackers with dietitians: Identifying design opportunities for food tracker customization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300822

[30] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. [n.d.]. Design guidelines for hands-free speech interaction. In *MobileHCI 2018*. 269–276. https://doi.org/10.1145/3236112.3236149

[31] Hyungik Oh, Jonathan Nguyen, Soundarya Soundararajan, and Ramesh Jain. 2018. Multimodal food journaling. In *Proceedings of the 3rd International Workshop on Multimedia for Personal Health and Health Care*. 39–47. https://doi.org/10.1145/3264996.3265000

[32] Ronaldo Parente, Ned Kock, and John Sonsini. 2004. An analysis of the implementation and impact of speech-recognition technology in the healthcare sector. *Perspectives in Health Information Management* 1 (2004). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2047322/

[33] Martin Pinquart and Silvia Sörensen. 2000. Influences of socioeconomic status, social network, and competence on subjective well-being in later life: a meta-analysis. *Psychology and aging* 15, 2 (2000), 187. https://doi.org/10.1037/0882-7974.15.2.187

[34] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. Accessibility came by accident: use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 459. http://doi.org/10.1145/3173574.3174033

[35] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 9, 3 (2002), 171–193. https://doi.org/10.1145/568513.568514

[36] Melanie Revilla, Mick P Couper, Oriol J Bosch, and Marc Asensio. 2020. Testing the use of voice input in a smartphone web survey. *Social Science Computer Review* 38, 2 (2020), 207–224. https://doi.org/10.1177/0894439318810715

[37] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23. https://doi.org/10.1145/3161187

[38] Kshitij Saxena, Robert Diamond, Reid F Conant, Terri H Mitchell, Guido Gallopyn, and Kristin E Yakimow. 2018. Provider adoption of speech recognition and its impact on satisfaction, documentation quality, efficiency, and cost in an inpatient EHR. *AMIA Summits on Translational Science Proceedings* 2018 (2018), 186. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961784/

[39] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in speech recognition*. Springer, 61–90. https://doi.org/10.1007/978-1-4419-5951-5_4

[40] Michael F Schober, Frederick G Conrad, Christopher Antoun, Patrick Ehlen, Stefanie Fail, Andrew L Hupp, Michael Johnston, Lucas Vickers, H Yanna Yan, and Chan Zhang. 2015. Precision and disclosure in text and voice interviews on smartphones. *PloS one* 10, 6 (2015), e0128337. https://doi.org/10.1371/journal.pone.0128337

[41] Katie Seaborn and Jacqueline Urakami. 2021. Measuring Voice UX Quantitatively: A Rapid Review. *arXiv preprint arXiv:2103.07108* (2021).

[42] Hyewon Suh, Nina Shahriaree, Eric B Hekler, and Julie A Kientz. 2016. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3988–3999.

[43] Stephen H Walsh. 2004. The clinician's perspective on electronic health records and how they can affect patient care. *Bmj* 328, 7449 (2004), 1184–1187. https://doi.org/10.1136/bmj.328.7449.1184

[44] Matthias Wenzel, Anja Perlich, Julia PA von Thienen, and Christoph Meinel. 2019. New ways of data entry in doctor-patient encounters. In *Design Thinking Research*. Springer, 159–177. https://doi.org/10.1007/978-3-319-97082-0_9

[45] Yixuan Zhang and Andrea G Parker. 2020. Eat4Thought: A Design of Food Journaling. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3334480.3383044