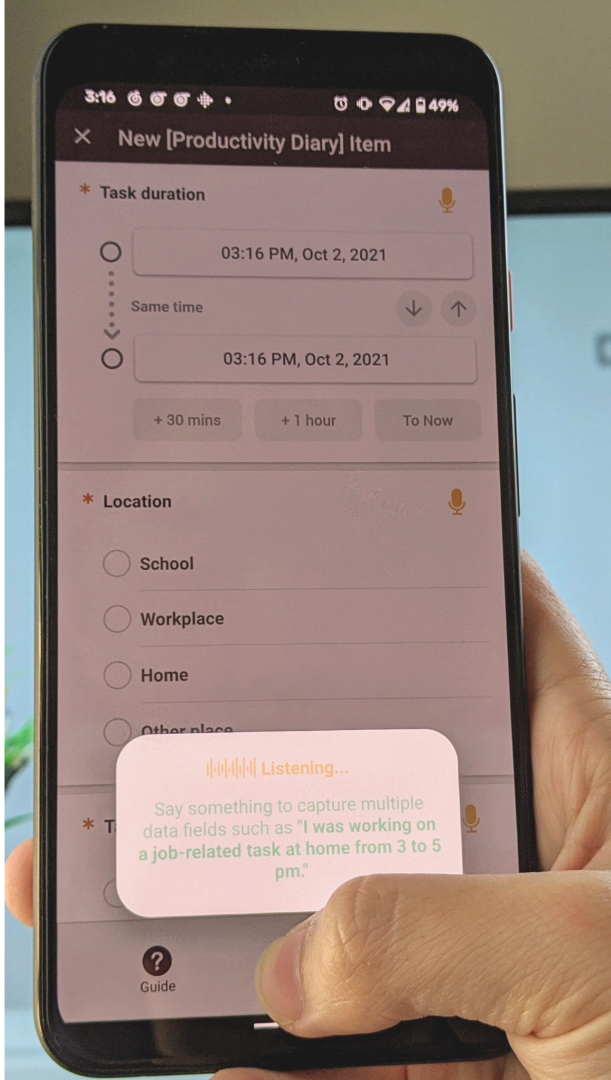
ACM Interactive Surfaces and Spaces ISS 2022
Sun 20 - Wed 23, November 2022 Wellington, NZ

NoteWordy: Investigating Touch and Speech Input on Smartphones for Personal Data Capture

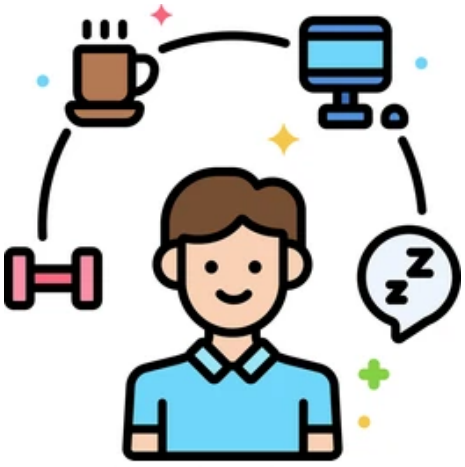
Yuhan Luo[1] Bongshin Lee[2]
Young-Ho Kim[3] Eun Kyoung Choe[4]

[1] City University of Hong Kong
[2] Microsoft Research
[3] Naver AI Lab, Republic of Korea
[4] College of Information Studies, University of Maryland

** Yuhan Luo and Young-Ho Kim conducted this research while at University of Maryland*



Self-tracking often involves capturing **multiple** data types

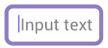


[Choe et al., 2014; Kim et al., 2017]

Time, duration

Activity type

Mood



Additional notes

The importance of manual tracking

- Collecting subjective/ contextual data
- Raising self-awareness

Icons created by Freepik - Flaticon

Traditional touch input & emerging speech input

+ Fast and easy
- Limited richness

+ Flexible input
- Heavy input burden

+ Low burden
+ Enhance data richness
- Difficult to edit?
- Environmental constraints?

[Luo et al., 2020; 2021; Korpusik et al., 2019; ModEat, Silva et al., 2021]

Limited research on how speech can support self-tracking

Audio recording without data processing [FoodScrap, Luo et al., 2021]

Extracting only single data type

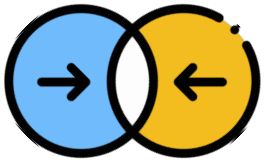
- Numbers [TandemTrack, Luo et al., 2021]
- Domain-specific items (e.g., food name and quantity) [Korpusik et al., 2019; ModEat, Silva et al., 2021]

Little understanding on how people use speech together with other input modalities

Integrating touch & speech input

+ Fast and easy
- Limited richness

+ Flexible input
- Heavy input burden



+ Low burden
+ Enhance richness
- Difficult to edit?
- Environmental constraints?

[Luo et al., 2020; 2021; Korpusik et al., 2019; ModEat, Silva et al., 2021]

Icons created by Freepik - Flaticon

Research Questions

RQ1. How do people use touch and speech input, individually or together, to capture different types of data for self-tracking purposes?

RQ2. How does the input modality affect people's data capture burden?

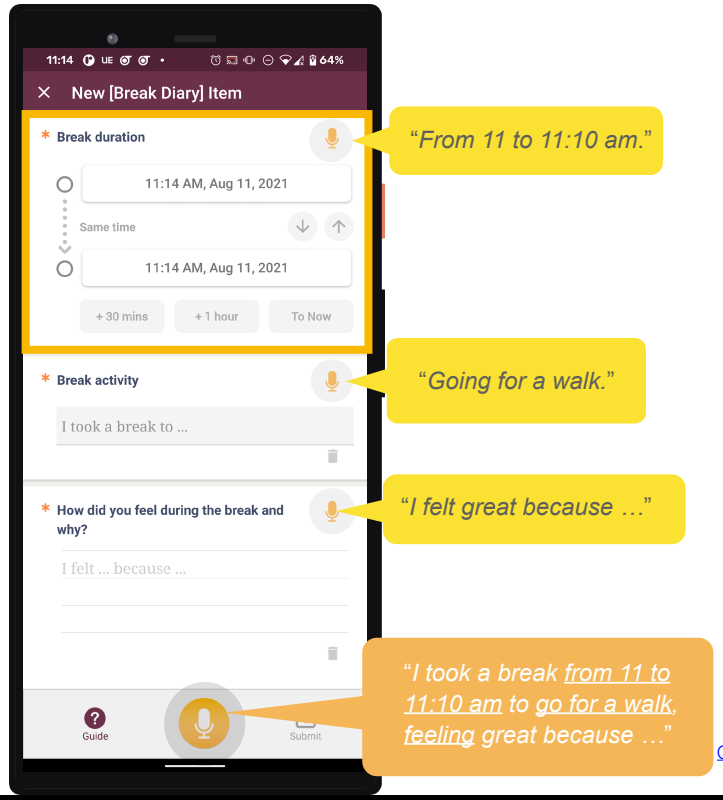
RQ3. How does the input modality affect the data richness of free-form text input?

NoteWordy: a multimodal self-tracking app

Touch input
Single/multiple taps or typing → one data field

Local speech (LS) input
One utterance → one data field

Global speech (GS) input
One utterance —> one/multiple data fields



[OmniTrack for Research](#) (Kim et al., 2019)

* One utterance: spoken input from the user at a time, could be a single word, an entire phrase, a sentence, or several sentences.

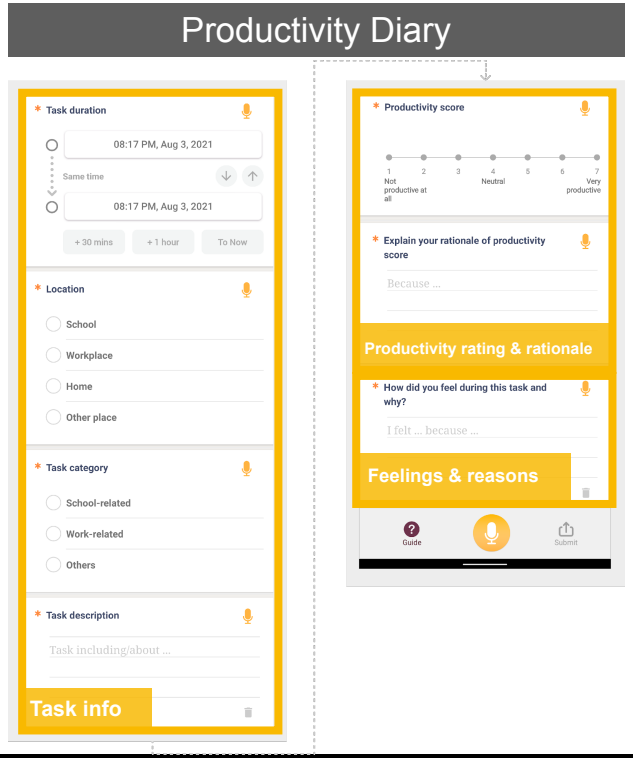
Research Context: **productivity tracking** for working graduate students

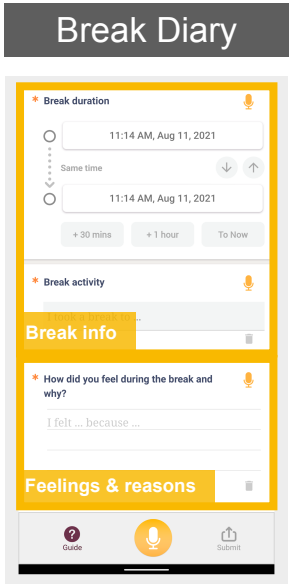
Image source: Center for The Analysis of PostSecondary readiness

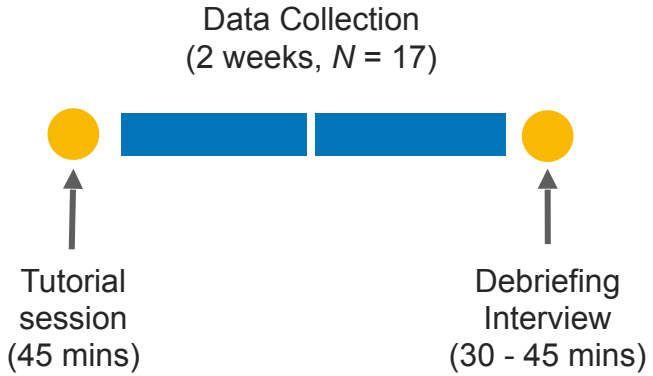
Productivity can be conceptualized in multiple dimensions corresponding to different data types (e.g., task duration, productivity level)
[Kim et al, 2019]

Working graduate students: juggle multiple tasks and struggle with maintaining a healthy balance between school and work
[Lee et al, 2017]

Data capture regimen and study procedure







Finding Highlights

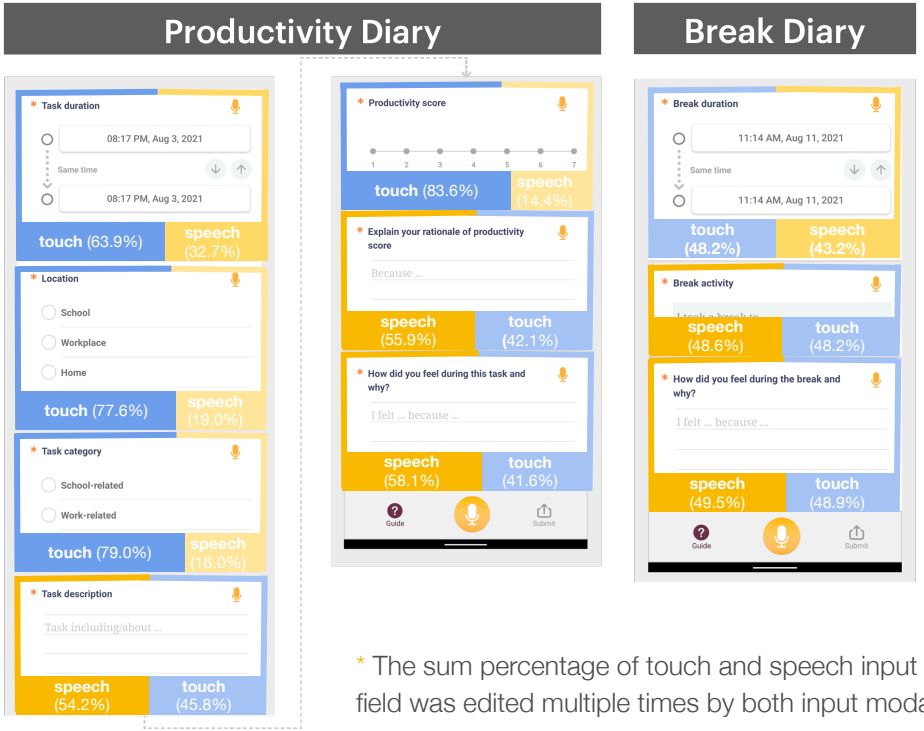
- Modality preferences
- Data Capture burden
- Data richness of free-form input

NoteWordy general usage

Diary	# of total entries	# of touch-only entries	# of speech-only entries	# of touch + speech entries
Productivity Diary	1032	429 (41.6%)	38 (3.7%)	565 (54.7%)
Break Diary	382	184 (48.2%)	131 (34.3%)	67 (17.5%)

* We use “speech-only input” to denote people using LS or GS input to enter their data, although it requires touching the speech button (i.e., the “push-to-talk” operation).

Modality choice by data type



Most **multiple choices**, and **Likert scale** were filled by touch input

Most **timespan** were filled by touch input, but also frequently filled by speech input

Text fields were commonly filled by both touch and speech input

* The sum percentage of touch and speech input < 100%, because there were a few (less than 5%) cases where a data field was edited multiple times by both input modalities

Using GS to capture multiple data fields

Start with GS to capture multiple data fields → adjust with touch or LS if needed

I was somewhat productive because …

"fast and intuitive" "more close to natural language"

I work on a work-related task at workplace

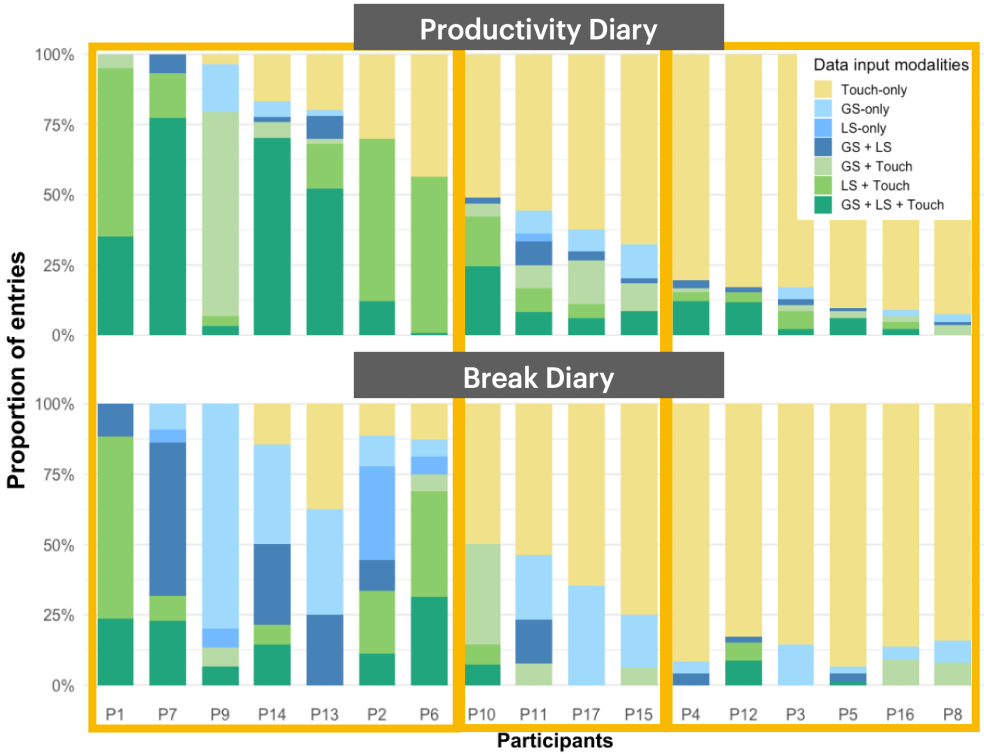
"awkward expression"

GS was used more frequently in Break Diary (43.2%) than Productivity Diary (22.1%)

I walked outside from 4 to 4:30 pm, feeling refreshed because the weather was nice

"short and straightforward"
"all the data fields on the screen"

Modality preferences vary across individuals



+ Convenience

+ Accuracy

- Privacy concerns

"*Not want share my productivity with colleagues*"

- Not a "social norm"

Worrying about oversharing

- More comfortable with touch input

"*Better at writing than speaking*" complicated thoughts

Average time spent (secs)

Diary	Avg	Touch-only entries	Speech-only entries	Touch + speech entries
Productivity Diary	143.7	175.9	115.9	121.1
Break Diary	78.4	86.7	65.5	81.0

Entries involving speech input took less time to complete than touch-only entries - speech input could help reduce entry completion time

The difference was significant in the **Productivity Diary** (b = -0.38, p = .004)

Speech recognition issues are the main hurdles

Number recognition

"*7 to 9*" → "*729*" (timespan recognition fail)

Misinterpretation

"*Moderately productive*" → Productive (original intention: *somewhat productive*)

Punctuation

"*It kept interpreting my pauses as periods when they should have been commas*" (P8)

Data richness in free-form input: task description

Generality	Specifics	Specifics with additional contexts
General description without concrete information	Specific about task details or the reasons of productivity rating / feelings	Specifics with contexts beyond the questions asked (e.g., task procedure, upcoming events)
"*Had a meeting*"	"*Met the team to discuss mockup design*"	"*I attended a UX meeting with other designers. We shared some case studies applying design thinking and talked to the BA team for next steps*"

Can input modality make a difference?

Task description: entries involving speech input tended to be specific *(OR = 3.79, p <.001)* and were more likely to include additional contexts *(OR = 3.0, p <.001)*

Productivity rationale: entries involving speech input tended to be specific *(OR = 2.16, p =.002)* and include additional contexts *(OR = 4.18, p <.001)*

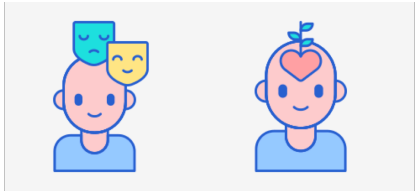
Feelings: entries involving speech input were more likely to include additional contexts *(OR = 2.12, p <.03)*

* *OR*: *Odds ratio*. An *OR* greater than 1 indicates that the condition or event is more likely to occur in that group

Implications

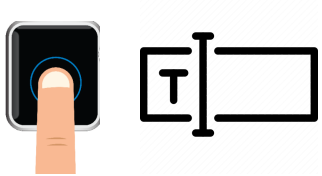
Integrating Touch & Speech to Capture Different Data Types

- Fast for free-form text
- Flexible time input
- Intuitive expression
- Enhance data richness



Capturing detailed contextual data (e.g., self-reported symptoms, mood, thoughts/feelings)

- Easier for single tap
- Quick editing support
- Account for privacy



Capturing structured/private data & supporting error correction

Supporting Efficient Multi-Data Capture With Speech Input

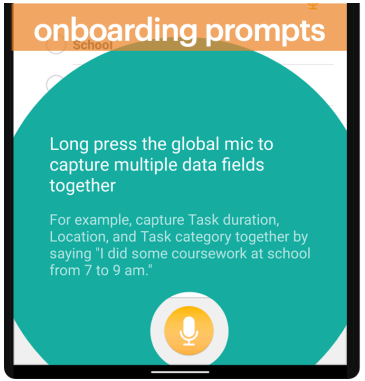
Arrange semantically-related data fields together

Time + Activity

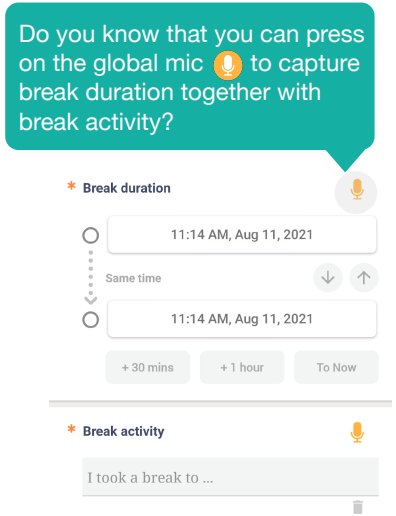
Time + Location

Rating + Rationale

Guided prompts to overcome unfamiliarity



The onboarding guide we provided in NoteWordy



Design opportunity: prompt people to try out GS when they press on LS

Adapting Speech Recognizers for Various Tracking Contexts

Context-agnostic speech recognizers are not fine-tuned for self-tracking data

E.g., Number recognition tends overlook the context (“*7 to 9*” —> “*729*”)

More research efforts are needed to contribute to the **contextualized speech data** from diverse self-tracking activities

- Date & time, duration
- Labels of Likert scale (e.g., stress level, sleep quality)
- Common units for daily activity (e.g., exercise repetitions)
- … …

Thanks!

Contributions

Design of **NoteWordy**, a multimodal self-tracking app integrating touch and speech input

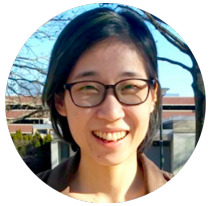
Empirical understanding of how speech works with touch input support people to capture different types of data for self-tracking purposes

Yuhan Luo, Assistant Professor, CS@CityU HK, yuhanluo@cityu.edu.hk

Acknowledgment


Bongshin Lee
Sr. Principal Researcher
Microsoft Research


Young-Ho Kim
Research Scientist
Naver AI


Eun Kyoung Choe
Associate Prof.
U of Maryland

Study participants and anonymous reviewers

NSF Awards #1753452 “Advancing Personal Informatics through Semi-Automated and Collaborative Tracking”