Older Adults' Think-Aloud Verbalizations and Speech Features for Identifying User Experience Problems

Mingming Fan
School of Information (iSchool)
Rochester Institute of Technology
Rochester, New York, USA
mingming.fan@rit.edu

Qiwen Zhao*
School of Information (iSchool)
Rochester Institute of Technology
Rochester, New York, USA
qz6021@rit.edu

Vinita Tibdewal
School of Information (iSchool)
Rochester Institute of Technology
Rochester, New York, USA
vt2173@rit.edu

ABSTRACT

Subtle patterns in users' think-aloud (TA) verbalizations and speech features are shown to be telltale signs of User Experience (UX) problems. However, such patterns were uncovered among young adults. Whether such patterns apply for older adults remains unknown. We conducted TA usability testing with older adults using physical and digital products. We analyzed their verbalizations, extracted speech features, identified UX problems, and uncovered the patterns that indicate UX problems. Our results show that when older adults encounter problems, their verbalizations tend to include observations (remarks), negations, question words and words with negative sentiments; and their voices tend to include high loudness, high pitch and high speech rate. We compare these subtle patterns with those of young adults uncovered in recent studies and discuss the implications of these patterns for the design of Human-AI collaborative UX analysis tools to better pinpoint UX problems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in accessibility**; **Usability testing**; **Empirical studies in HCI**.

KEYWORDS

older adults, elderly, seniors, think-aloud, verbalization, speech features, UX problems, usability testing, remote usability testing, AI-assisted UX analysis, human-AI collaboration for UX analysis

ACM Reference Format:
Mingming Fan, Qiwen Zhao, and Vinita Tibdewal. 2021. Older Adults' Think-Aloud Verbalizations and Speech Features for Identifying User Experience Problems. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445680

1 INTRODUCTION

Think-aloud (TA) protocols are widely used by user experience (UX) practitioners to elicit users' thought processes, that are otherwise unavailable to observers, when they interact with a user interface, to uncover and understand UX problems [18, 33]. There has been an increasing interest in analyzing users' think-aloud verbalizations at micro-levels, such as categorizing verbalizations into different categories [10, 14, 17, 24, 43], to better understand their experiences.

Recent research found that subtle patterns in users' verbalizations and speech features are telltale signs of UX problems [17]. For example, users tend to verbalize a particular type of utterances, use words with negative sentiment, slow down their speech, or raise their tone when they encounter problems [17]. As analyzing think-aloud usability testing session is time-consuming [18, 33], researchers leveraged such patterns to build artificial intelligence (AI) models to detect UX problems automatically [16]. To overcome the limitation of AI, researchers recently began to explore AI-Assisted UX analysis methods to analyze large amounts of think-aloud usability test sessions more efficiently [19].

Despite the promise and potential applications of such patterns in users' verbalizations and speech features, these patterns were uncovered among young adults (aged 19-26) [17]. It is unclear whether and to what extent such patterns still exist among other age groups, such as older adults. As research has shown differences in task performance of think-aloud sessions between older and young adults [35, 36], such differences might be reflected in their verbalizations and thereby in subtle patterns indicating UX problems. Further, if subtle patterns do exist for older adults, identifying and quantifying such patterns would inform the design of automatic and human-in-the-loop analysis methods to better identify UX problems that older adults encounter. Therefore, in this research, we seek to understand older adults' think-aloud verbalizations and speech features and how they might indicate UX problems.

We first conducted think-aloud usability testing with older adults using both physical and digital products. We then categorized participants' verbalizations using categorization strategies proposed in the literature [10, 14, 17]. We further annotated sentiments and extracted speech features (e.g., loudness, pitch, and speech rate) for each segment. Next, we computed how well verbalization categories, sentiments and speech features were indicative of UX problems in terms of precision, recall, and F-measure. Moreover, we extracted and compared most frequently verbalized words when participants did and did not encounter problems. Finally, we discussed our findings in the context of prior work among young adults (e.g., [17, 24, 44]) and the implications. In sum, we make the following contributions:

- Identification and quantification of the subtle patterns in verbalization *categories*, *sentiments*, and *speech features (e.g., loudness, pitch, and speech rate)* that indicate UX problems for *older adults*.

*First student author

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445680

- The implications of these patterns for future UX analysis methods (e.g., automatic methods, human-AI collaboration tools).

2 BACKGROUND AND RELATED WORK

2.1 Think-Aloud Protocols

Think-aloud (TA) protocols were initially developed in psychology to study people's thought processes when they solve problems [15] and later introduced into Human-Computer Interaction to identify UX problems [29]. A recent survey study with 197 UX practitioners from various sized companies in different geographic locations showed that majority of them (86%) used TA protocols [18], which confirmed the findings of an earlier international survey study [33].

There are two categories of TA protocols: *concurrent* and *retrospective* protocols. In concurrent TA protocols (CTAs), users verbalize their thought processes when working on the task at the same time; in retrospective TA protocols (RTAs), users verbalize their thought processes, after completing the task, when watching the recording of the session. CTAs are more widely adopted than RTAs in practice [18]. Within CTAs, there are three variations depending on the types of interventions from the moderator. In the *classic* CTA, the moderator only reminds users to keep talking if they fall into silence for a period [15]. In the *speech-communication* CTA, the moderator also uses speech tokens (e.g., "and then...") to elicit verbalizations from users [5]. In the *interactive* CTA, the moderator actively probes users by asking questions [12, 38]. Researchers have studied differences in these protocols and suggested to use the classic CTA over the other two CTAs because it is as effective in identifying UX problems as the other ones and least likely threatens the validity of users' verbalizations [1, 2]. Therefore, in this research, we adopted the classic CTA to conduct think-aloud usability test sessions. What's more, this would allow for comparing our findings with prior work that uncover subtle verbalization patterns with young adults as their study also used classic CTA [17].

2.2 Think-Aloud Verbalizations

Previous research scrutinized think-aloud verbalizations at micro-levels by categorizing them into different *verbalization categories*. In an early work, Cooke categorized users' verbalizations in CTA into five categories: Reading, Procedure, Observation, Explanation, and Other [10]. Later, Elling et al. analyzed their participants' verbalizations in CTA and found the same set of verbalization categories [14]. Other researchers unpacked more detailed categories than Cooke's five-category scheme. For example, Hertzum et al. further divided the Observation category into four sub-categories, which were system observation, redesign proposal, domain knowledge, and user experience [24]. Zhao et al. identified ten more specific categories, which could be mapped in to Cooke's five-category scheme [43]. Recently, Fan et al. studied young adults' verbalizations in CTA sessions and also adopted Cooke's five-category scheme [17]. As Cooke's five-category scheme had been widely used or extended by prior studies [14, 16, 17, 24, 26, 43], we also adopted Cook's five-category scheme when categorizing older adults' verbalizations.

TAs were used to study older adults' thought process to understand their experiences with different software tools (e.g., [6, 9, 27, 30, 32]). For example, TAs were used to understand how older adults search health related information [27] or search online to interpret symptoms of illness [32]. TAs were also used to evaluate the usability of a health and wellness technology tool with older adults [9] or to understand older adults' motivations and challenges participating in crowd work [6].

In this work, we took a first step to understand older adults' verbalizations at micro-levels by uncovering different verbalization categories, sentiments of verbalizations, speech features extracted from TA audios, and even the most frequently verbalized words. With these micro-level analyses, we sought to uncover how these features indicate UX problems.

2.3 Think-Aloud Verbalizations and UX Problems

People's utterances (i.e., verbalizations) and speech features (i.e. pitch, loudness, speech rate,and sentiment) have shown to be able to reveal how they feel [37], whether they are experiencing cognitive overload [13, 40], and whether they are confident in their tasks [13, 28, 37]. Recently, Fan et al. studied young adults' verbalizations and speech features during CTA usability test [17] and found that when encountering a UX problem, users tended to verbalize utterances of the Observation category or utterances associated with negative sentiment, raise their speech tone, or slow down their speech rate [17]. Analyzing TA usability test sessions is often time-consuming because it entails reviewing session videos, often repeatedly, and scrutinizing participants' verbalizations to pinpoint UX problems [18, 33]. Inspired by the subtle patterns [17], researchers started to leverage the patterns to build artificial intelligence (AI) models to detect UX problems automatically [16]. To alleviate the limitations of AI, researchers recently began to explore AI-Assisted human-in-the-loop methods, such as visualizing AI's predicted problems, to help UX evaluators more effectively analyze and derive insights from large amounts of TA usability test sessions [19]. Although such verbalization and speech patterns have many potential applications, they were uncovered with young adults [17]. It is still unknown whether and how patterns in older adults' verbalizations and speech features are indicative of UX problems. This has motivated us to answer the following Research Questions (**RQs**):

- *What are subtle patterns in older adults' verbalizations and speech features that are indicative of UX problems?*
- *How are the subtle patterns compared to those uncovered with young adults? What are the implications of the subtle patterns for UX analysis methods?*

3 METHOD

We present the details of our IRB-approved study in this section.

3.1 Participants

We recruited participants via advertisements posted in local senior community centers and word-of-mouth. In the end, ten participants (7 females and 3 males) completed the study, who aged between 62 and 85 ($M = 75, SD = 7$). The quick spread of COVID prevented us from conducting more in-person studies. All participants were native English speakers and had no physical or cognitive impairments that prevented them from interacting with the test products independently in the study.

3.2 Test Products and Tasks

We chose three different types of products for the CTA usability testing to evaluate the potential effect of the test products. We included both physical and digital products to increase their representativeness. For the physical device, we chose the same coffee machine as the one used in a recent study that uncovered the subtle patterns among young adults [17]. Figure 1 shows the coffee machine used by a participant in the study. For digital products, we chose a pet adoption website and a food delivery mobile app. Figure 2 shows the key pages of the website, and Figure 3 shows the mobile app used by a participant.

Figure 1: The coffee machine setup in the study

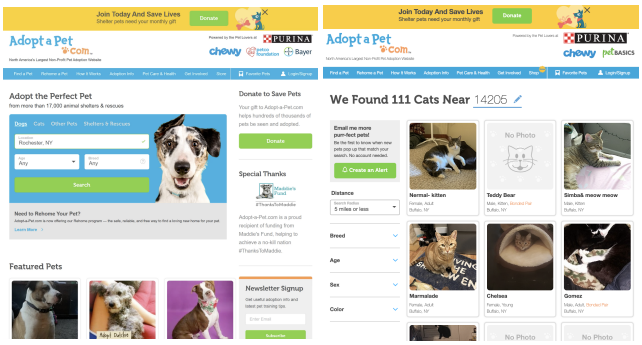

Figure 2: The key sub-pages of the test website

We selected the website and the mobile app that cover common computer usage (e.g., browsing internet and using mobile apps) and represent products older adults may use on a regular or occasional basis. Furthermore, these products also contained UX problems as identified through heuristic evaluation conducted by the research team. All the participants had not used these particular products prior to the study.

Table 1 shows the test products and the tasks. We replicated the tasks used for the coffee machine in a recent study [17] for comparison. The coffee machine could be programmed to make coffee at a set time, and thus the task included a time setting step. The tasks for the website and the mobile app were related to their main functions and contained UX problems, which were identified through heuristic evaluation conducted by the research team. For the food delivery mobile app, there were two subtasks. Participants first worked on the "pick-up" subtask (the first two requirements in Table 1). Then, they worked on the "delivery" subtask (the las requirement in Table 1). The order of the test products was randomized to minimize the potential order effect.

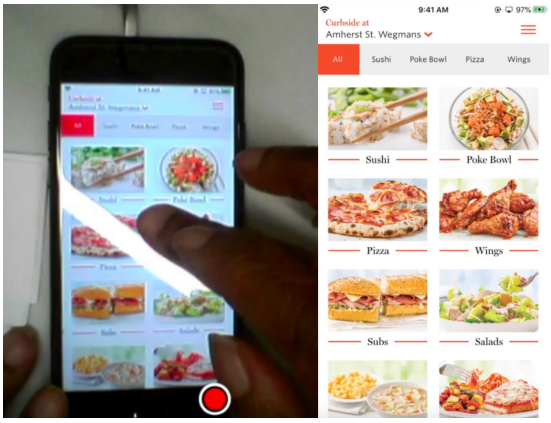

Figure 3: A participant's hand interacting with the food delivery mobile app and the home screen of the mobile app.

3.3 Procedure

We first introduced the classic concurrent think-aloud (CTA) protocol to participants and played a short online video tutorial on CTA [22]. Next, we asked participants to practice think-aloud by offering them an alarm clock and asking them to set up an alarm. Then, participants started to perform three formal think-aloud sessions, in which they worked on tasks with three products (see Table 1). During the study, the moderator followed Ericsson and Simon's guidelines for classic CTA [15] and did not interact with participants, except for reminding them to keep talking if they fell into silence for a while. The moderator did not offer any help on the tasks except for asking participants to move forward if they were stuck on an issue for more than three minutes. To avoid fatigue, the total length of the study was around an hour, which included the time for introduction of CTA and TA practice, the time for completing tasks, and rests between tasks. As a result, the maximum time for each task was about 10 minutes. If participants did not complete the task within the time limit, the moderator would stop the task and ask participants to take a rest and then move on to the next task. Each participant was compensated with $25 cash.

All sessions were audio and video recorded. To better capture the participant's audio, we attached a clip-on microphone to the participant's collar. For recording coffee machine sessions, we used GoPro, which attached to the participant's chest using a chest mount, to capture the participant's hand motions, and we also used a camera to capture the overall view of the participant and the coffee machine (see Figure 1). For recording website sessions, we used a screen recording application. For recording mobile app sessions, we used Mr. Tappy, a camera and device hold kit [41], to record the participant's hand operations and a screen recording application to record the mobile app's screen.

Table 1: Products and Tasks used in the concurrent think-aloud usability testing of our study.

Products	Tasks
Coffee Machine	Program the coffee machine to make two cups of strong flavor drip coffee at 7:30 in the morning
Pet Adoption Website	Please try to find a cat that best matches the following requirements: 1) The cat owner needs to be from Buffalo, NY 14205, so that you can meet the cat more easily; 2) You already have one dog at your house, so the new cat needs to be good with dogs; 3) Your 5-year old granddaughter will visit you every weekend, so you would like the new cat to be kid-friendly; 4) The application fee should be no more than $100.
Food Delivery Mobile App	You will hold a party in your house and want to order some food in advance using the food delivery mobile app and collect them at the store in person: 1) Find the Wegmans store on the Amherst St.; 2) You will have 10 guests visiting your house. Your budget is $100. You would like to buy: a. 10 bottles of classic Coke and 10 bottles of Sprite; b. Some full sheet pizzas, with any topping you like. You can buy as many as you want, but they should be under your budget; 3) You realize that you won't do grocery shopping this week, so you cannot pick up the food in store. You want them to deliver the food to your home instead.

4 ANALYSES

We present the details about how we categorized participants' verbalizations, labeled sentiments, extracted speech features, identified UX problems, and quantified the patterns in users' verbalizations and speech features that are indicative of UX problems.

4.1 Categorizing Verbalizations into Categories

First, two researchers reviewed each recorded test video and broke it into small segments based on the pauses between users' verbalizations and the semantics of verbalizations. For each segment, the researchers then manually transcribed users' verbalizations.

For each segment, two researchers independently assigned a category label using an updated version of Cooke's verbalization categories (See Section 2.2). Table 2 shows the definitions of the verbalization categories and examples from our study. After independent categorization, the two researchers discussed to address the disagreements. If they could not reach a consensus, a third researcher joined the discussion to consolidate the category labels.

4.2 Annotating Verbalizations with Sentiments

The researchers followed a similar process to assign one of the three sentiment labels to each segment: *negative*, *positive* and *neutral* sentiments. A segment was labeled with a negative sentiment if it contained: a) negative expressions, such as *"This is way too complicated"*, b) confusions or frustrations about the product, such as *"Why do they not have the instructions on it?"* Similarly, a segment was labeled with a positive sentiment if it contained: a) positive expressions, such as *"Oh, that's cute"*, b) words indicating successful problem-solving, such as *"Got it."*. The rest of the segments were labelled with neutral sentiments.

4.3 Extracting Speech Features

For each segment in a recorded TA session, we computed *loudness* and *pitch* (i.e., fundamental frequency *F*0) from the corresponding audio at the sampling rate of 100 Hz using praatUtil library [23], which interfaces with speech process toolkit Praat 6.0.13 [4]. We set the frequency range to be 50-400 Hz to cover both typical male and female frequencies and also to filter out non-voice sounds. Moreover, we computed the *speech rate* of a segment by dividing the number of words spoken in the segment by its duration.

Because the sampling rate of extracting speech features was 100 Hz, there were 100 values extracted per second for each speech feature (e.g., loudness, pitch). As accidental environmental noises can cause abnormally high or low values, our algorithm considered a segment having high/low speech features only if the percentage of high/low values in the segment exceed a threshold to reduce noise effects. We tested a range of threshold values and found that 8% worked the best on our data. In other words, if more than 8% of the loudness (pitch) values of a segment were two standard deviations higher or lower than the average loudness (pitch) of the entire session, this segment would be labeled as having *high* or *low* loudness (pitch).

4.4 Identifying UX Problems

Two evaluators identified the UX problems appeared in the think-aloud sessions by analyzing 1) transcripts of the audio recordings 2) video recordings, and 3) field notes from the sessions. They labelled each segment whether the participant encountered a usability problem. Then, the two evaluators discussed when there was a disagreement and the third evaluator would join in when needed.

4.5 Quantifying How Subtle Patterns Indicate UX Problems

To compute the correlation between a verbalization category with UX problems, we computed the *precision*, *recall*, and *F-measure* of each verbalization category for identifying UX problems. Precision, recall, and F-measure are commonly used to measure the prediction power of a classifier. We used them to measure the prediction power of verbalization categories, sentiments, and speech features for locating UX problems. These measures had been used to quantify the connection between verbalization categories and speech features and UX problems in recent research [16, 17].

Let's denote i to be the verbalization category i. $N_{segment}(i)$: the number of verbalization segments that are labeled as the category i. Thus, $\sum_i N_{segment}(i)$ means the total number of segments of all verbalization categories. $N_{problem}(i)$: the number of verbalization segments that are labeled as the category i and are also associated

Table 2: The definition of the verbalization categories and corresponding examples.

Categories	Definition	Examples
Reading	Reading words, phrases, or sentences from the test product	*"Amherst street. Buffalo, New York, USA. OK."* -P10
Procedure	Describing current or future activities	*"I'm gonna add tomato sauce. I'm gonna add shredded mozzarella...So I'm gonna press the button for the toppings."* -P08
Observation	Making an observation or a remark about the test product or themselves	*"There is no Amherst. Not listed here. Hum."* -P04 *"It [referring to the search engine on the pet adoption website] is dumb. I don't even say it's stupid. It's just dumb."*-P01
Explanation	Explaining the reason or providing the motivation for their actions or behaviors	*"Since I don't know the machine, I'm going first to the manual."* -P09
Others	Verbalizations that do not fit into one of the above four categories	*"Last time I made coffee, it was not good. I forgot to put the cup underneath."* -P05

with a UX problem. Thus, $\sum_i N_{problem}(i)$ means the total number of problems associated with all verbalization segments. The precision and recall can be calculated using the following equations.

$$Precision = \frac{N_{problem}(i)}{\sum_i N_{segment}(i)}, Recall = \frac{N_{problem}(i)}{\sum_i N_{problem}(i)}$$

We adopted a standard equation of *F-measure*, which combines precision and recall into a comprehensive measure: $F{-}measure = \frac{2*Precision*Recall}{Precision+Recall}$. It is worth noting that precision, recall, and F-measure take the total number of verbalization segments into account, and thus are not affected by the total amount of verbalizations or the length of the usability test video. What's more, while these measures are usually used to quantify the performance of a machine learning (ML) classifier, they can be used outside of ML and are independent of the amount of data. As a result, although our data set was relatively small, these measures were still suitable measures to quantify the relative correlations between various think-aloud verbalization and speech features and the UX problems.

Intuitively, if a verbalization category has a higher precision, it means that UX evaluators would have a higher chance to find a UX problem when examining a segment of the category. If a verbalization category has a higher recall, it means that UX evaluators would be able to find a higher percentage of UX problems by examining all segments of the category. If a verbalization category has a higher F-measure, it means that UX evaluators would have an overall better chance to locate a UX problem by checking the segments of the category.

In addition to verbalization categories, we computed these three measures to quantify how the following features indicate UX problems: sentiment, loudness, pitch, and speech rate.

4.6 The Most Frequently Verbalized Words

To better understand what participants verbalized when they encountered or did not encounter problems, we extracted the most frequently verbalized words from their verbalizations. We first used the *word_tokenize* function in the natural language processing toolkit (NLTK) [31] to extract the list of words appeared in the transcripts of all think-aloud sessions. We then removed numbers from the list and converted the words into lowercase. Next, we removed punctuation marks and stop words (i.e., a set of commonly used words in a language) from the list. Example stop words include all forms of pronouns (e.g., I, me, my, mine, myself), prepositions (e.g., in, on, at), and all forms of auxiliary verbs (e.g., be, do, have). Finally, we used the *FreqDist* function in the NLTK to compute the frequency of all the remaining words and extracted the top 25 most frequently appeared words when participants encountered or did not encounter problems.

5 RESULTS

5.1 UX Problems

The average number of UX problems encountered by the participants for the website, mobile app, and coffee machine were 23 (SD = 12), 27 (SD = 11) and 21 (SD = 8). We describe the UX problems with examples and usability heuristics violated in appendices.

5.2 Verbalization *Categories* and UX Problems

5.2.1 Verbalization Category Proportions. Table 3 shows the number and percentage of segments that were of each verbalization category. The most frequent to the least frequent verbalization category was as follows: Observation, Reading, Procedure, Explanation, and Others. This trend was overall consistent for different products.

Table 3: The number and percentage of segments per *verbalization category* for each product and all products together.

	Coffee Machine	Website	Mobile App	All Products
Reading	110 (18.71%)	58 (18.13%)	114 (20.28%)	282 (19.18%)
Procedure	86 (14.63%)	31 (9.69%)	60 (10.68%)	177 (12.04%)
Observation	356 (60.54%)	219 (68.44%)	343 (61.03%)	918 (62.45%)
Explanation	20 (3.40%)	9 (2.81%)	21 (3.74%)	50 (3.40%)
Others	16 (2.72%)	3 (0.94%)	24 (4.27%)	43 (2.93%)

Table 4: The precision, recall, and F-measure of five verbalization categories for identifying UX problems for each test product.

	Coffee Machine			Website			Mobile App		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Observation	0.51	0.91	0.66	0.55	0.87	0.68	0.49	0.71	0.58
Procedure	0.09	0.04	0.06	0.10	0.02	0.04	0.28	0.07	0.11
Reading	0.02	0.01	0.01	0.16	0.06	0.09	0.31	0.15	0.20
Explanation	0.20	0.02	0.04	0.56	0.04	0.07	0.33	0.03	0.05
Others	0.31	0.02	0.05	0.33	0.01	0.01	0.38	0.04	0.07

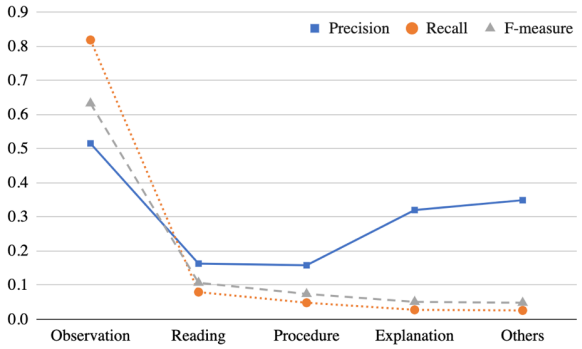

Figure 4: Precision, recall, and F-measure of each verbalization category for identifying UX problems for all products.

5.2.2 Correlations between Verbalization Categories and UX Problems. Figure 4 shows the precision, recall, and F-measure of each verbalization category for identifying UX problems for all products together. We further calculated precision, recall, and F-measure of each verbalization category for identifying UX problems for three products respectively. Table 4 shows the result. The trend between verbalization categories and problems was consistent for each product.

The average F-measure of the five categories (observation, procedure, reading, explanation, others) across all products were 0.64, 0.07, 0.10, 0.05, and 0.04 respectively (based on Table 4). Observation was the most indicative of UX problems, and its prediction power, based on F-measure, was more than 6 times of any other verbalization categories (0.64 was more than six times bigger than the rest of the numbers). Figure 4 also shows that Explanation and Others had lower recall and F-measure than Reading and Procedure. It means that UX evaluators would find fewer UX problems when checking Explanation or Others segments than checking Reading or Procedure segments. However, Explanation and Others had relatively higher precision than Reading and Procedure. It means that UX evaluators would have a better chance to locate a UX problem by randomly checking an Explanation or Others segment than a Reading or Procedure segment. Next, we present examples of TA verbalizations to illustrate how the verbalizations of each category indicated UX problems.

Observation: Observation category contained remarks that participants made about the user interface (UI) or themselves (Table 2). Remarks about their observations of the UI could suggest problems. In the coffee machine session, P4 pressed the "1-5 cups" button since she was tasked to make two cups of coffee. When seeing no feedback, P4 verbalized, "*Why it didn't light up?*", "*Nothing's happening*". In the mobile app session, P2 could not find the shopping cart to edit her order, so she verbalized, "*Oh wait. Where is my cart? Where is my list of my cart?*" Further, remarks related to their experiences could also suggest problems. After trying to program the coffee machine but failing every time, P6 verbalized, "*Oh gosh. This is aggravating.*"

Procedure: Procedure category contained verbalizations of their current actions or actions that they are about to take. In the mobile app session, P8 selected a wrong store to place the order due to the poor size of the tabs in the app. He had to re-enter the zip code and search the store again: "*Ah. I'm gonna go back and put in my zip again...*"

Reading: Reading category contained verbalizations of reading instructions or information from the UI. When participants read for a long time or kept repeating some content, it was often a signal of them experiencing problems. In the website session, P10 didn't know how to use the search engine to filter the cats she wanted. Instead, she looked at the navigation bar repeatedly, hoping to use the navigation bar to search the cat. She kept reading the items on the navigation bar, "*Find a cat. Rehome a cat. Find a cat...*"

Explanation: Explanation category contained verbalizations of their motivation and intention. Product design could cause users to form incorrect motivation. In the coffee machine session, P7 was confused about the two coffee powder measuring cups and verbalized, "*It looks about the same size, so I better check the directions.*"

Others: Others contained verbalizations that did not fit directly into the above four categories. Such verbalizations were often not directly related to the task at hand. In the website session, P1 indicated that computers could not sympathize with people by verbalizing, "*It's all based on magic. Computers can't sympathize. Can they? That's AI all about.*" While it was not directly related to the task, it still provided us a side-note about his (unmet) expectation for the website.

5.3 Verbalization *Sentiments* and UX Problems

5.3.1 Sentiment Proportions. Table 5 shows the number and percentage of the segments labeled with each type of sentiment for each and all products together. The most frequent to the least frequent sentiment type for all products was as follows: neutral, negative, and positive. Moreover, there were roughly twice as many segments with neutral sentiments as with negative sentiments. This trend was consistent for three products.

Table 5: The number and percentage of segments per *sentiment* type for each test product and all products together.

	Coffee Machine	Website	Mobile App	All Products
Neutral sentiments	356 (60.54%)	182 (56.88%)	359 (63.88%)	897 (61.02%)
Negative sentiments	175 (29.76%)	107 (33.44%)	165 (29.36%)	447 (30.41%)
Positive sentiments	57 (9.69%)	31 (9.69%)	38 (6.76%)	126 (8.57%)

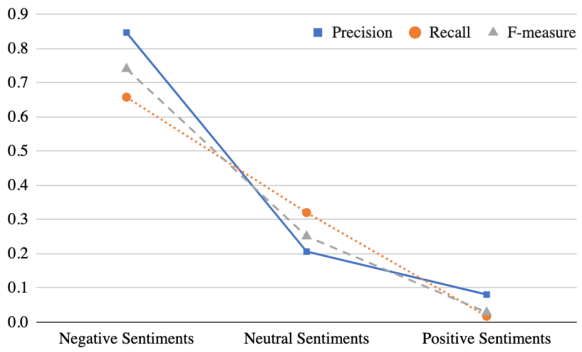

Figure 5: Precision, recall, and F-measure of each type of sentiments for identifying UX problems for all products.

5.3.2 Correlations between Sentiments and UX Problems. Figure 5 shows the precision, recall, and F-measure of each type of sentiments for identifying UX problems. Negative sentiments were the most indicative of UX problems with the average F-measure of 0.74, followed by neutral and positive sentiments. Next, we present verbalization examples of three types of sentiments to illustrate how they indicated UX problems.

Negative sentiments: Participants' negative experiences with products often resulted in their verbalizations of negative sentiments. When P1 failed to program the coffee machine, he verbalized, "*This is goofy. Too complicated.*", and "*I am so frustrated.*"

Neutral sentiments: Although it is not as often as negative sentiments, participants' frustrations could also be expressed in verbalizations with neutral sentiments. When P1 encountered some problems typing on the website, he verbalized, "*I know how to type, Just don't do it all the time.*"

Positive sentiments: Interestingly, participants' verbalizations with positive sentiments could also occasionally suggested problems. One example was sarcasm. In the coffee machine session, P3 did not understand what the "auto" button means, but he verbalized, "*Uhmmm, that is interesting.*"

We further computed precision, recall, and F-measure of each type of sentiment for identifying UX problems for three products respectively. Table 6 shows the result. The trend between different types of sentiments and problems was consistent for each product.

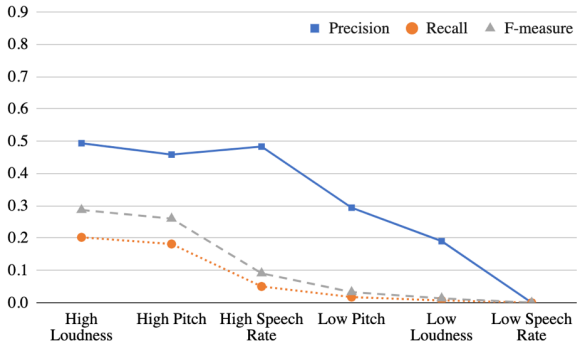

Figure 6: Precision, recall, and F-measure of each speech feature for identifying UX problems for all products.

5.4 *Speech Features* and UX Problems

5.4.1 Speech feature proportions. Table 7 shows the number and percentage of segments per speech feature for each product and all products together. There were only a small number of segments associated with either high or low speech features. Specifically, segments with high loudness, pitch and speech rate were only about 16.12%, 15.58%, and 4.08%; segments with low loudness, pitch and speech rate were even more rare (all below 4%).

5.4.2 Correlations between speech features and UX Problems. Figure 6 shows the precision, recall, and F-measure of six speech features (i.e., high/low loudness/pitch/speech rate) in locating UX problems. Overall, these high and low speech features had much lower recall and F-measure than the Observation category or Negative sentiments. Nonetheless, high speech features (i.e., loudness, pitch and speech rate) had reasonably well precision (close to 0.5). What's more, these high speech features were more indicative of UX problems than low speech features. Next, we present verbalization examples with these speech features to illustrate how they indicated UX problems.

High Loudness: Participants raised their volume when they were extremely agitated. For example, in the coffee machine session, at several points P6 encountered difficulties that she could not solve, she became agitated and raised the volume of her speech. She said, *"That didn't work well.", "Can I quit?"*

High Pitch: Participants also raise their pitch when something out of their expectations happened. For example, in the coffee machine session, P6 raised her pitch and laughed, when she was not able to program the coffee machine by pressing the hour button on the panel. She said, *"[Keep pressing the hour button] Didn't do anything...Alright. Hahaha. Oh, this is horrible. Hahaha"*

High speech rate: We noted that participants spoke slowly when thinking aloud. However, when they increased their speech rate, it was likely because they temporarily put aside the task to make comments on the products. For example, in the mobile app session, P3 struggled with using the app to order food, so he paused the task and asked whether he could call the store to place the order instead of using the app: *"That would be easier. Just give them a ring*

Table 6: The precision, recall, and F-measure of three types of sentiments for identifying UX problems for each test product.

	Coffee Machine			Website			Mobile App		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Negative	0.88	0.76	0.82	0.84	0.65	0.73	0.81	0.57	0.67
Neutral	0.13	0.22	0.16	0.26	0.34	0.29	0.26	0.39	0.25
Positive	0.02	0.01	0.02	0.06	0.01	0.02	0.18	0.03	0.01

Table 7: The number and percentage of segments per *speech feature* for each test product and all products together.

	Coffee Machine	Website	Mobile App	All Products
High Loudness	111 (18.88%)	48 (15.00%)	78 (13.88%)	237 (16.12%)
High Pitch	105 (17.86%)	48 (15.00%)	76 (13.52%)	229 (15.58%)
High speech rate	26 (4.42%)	15 (4.69%)	19 (3.38%)	60 (4.08%)
Low Loudness	0 (0%)	2 (0.63%)	19 (3.38%)	21 (1.43%)
Low Pitch	2 (0.34%)	11 (3.44%)	21 (3.74%)	34 (2.31%)
Low speech rate	0 (0%)	0 (0%)	8 (1.42%)	8 (0.54%)

and tell them what I want, right? And I will be there in an hour to pick it up."

Low loudness, pitch, and speech rate: Participants seldom verbalized in low speech features (Table 7). Consequently, such features were also unlikely related to problems (Figure 6). Nonetheless, there were few instances where such low speech features indicated problems. One representative example was when participants fell into mumbling. In the Mobile App session, P7 was struggling to type in the address of the store and started to speak at a reduced volume and lower pitch and mumbled, "*601 [Amherst Street]. I don't do this [referring to using the keyboard on the mobile phone].*"

We further computed precision, recall, and F-measure of each type of sentiment for identifying UX problems for three products respectively. Table 8 shows the result. The general trend between speech features and problems was consistent for each product.

The list of the top 25 frequently verbalized words when participants encountered problems is shown in Figure 7. The list included *negations* (e.g., "n't", "not", and "no"), *question words* (e.g., "what", "how", "where", and "why"), *words with negative sentiments* (e.g., "back", "but"), *filler words* (e.g., "oh", "OK", "alright", and "well"), *common verbs* (e.g., "find", "got", "na" as part of "wanna" or "gonna","gon" as part of "gonna", and "think"), and *task-related nouns* (e.g., "pizza", "button", and "thing").

Similarly, the list of the top 25 frequently verbalized words when participants did not encounter problems are shown in Figure 8. The list included a significant number of *task-related nouns* (e.g., "instructions", "two", "pizza", "coffee", "cart", "button", "cups", and "cat"), *common and task-related verbs* (e.g., "reading", "na" as part of "wanna" or "gonna", "gon" as part of "gonna", "need", "got", "put", "add"), *negation* (e.g., "n't"), *filler words* (e.g., "OK", "alright", "oh"), *words with positive sentiment* (e.g., "good", "right"), and also *question words* (e.g., "what", "how").

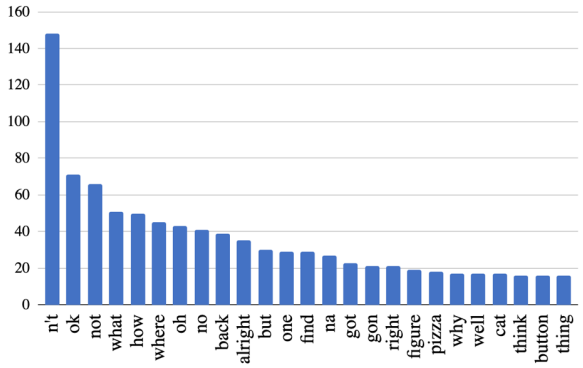

Figure 7: The top 25 frequently verbalized words when participants encountered problems.

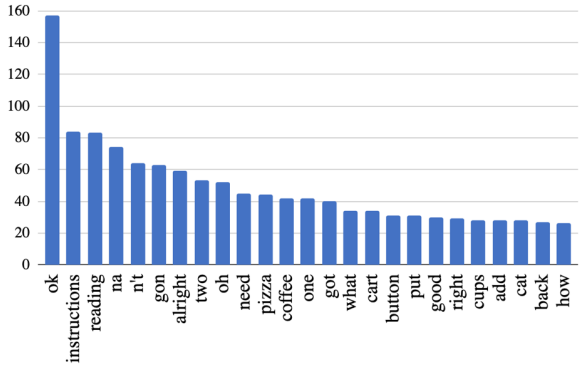

Figure 8: The top 25 frequently verbalized words when participants did not encounter problems.

6 DISCUSSION

We first present the key takeaways in Section 6.1 and then elaborate on them in the subsequent Sections 6.2, 6.3, 6.4, and 6.5. Finally, we discuss the design implications in Section 6.6.

Table 8: The precision, recall, and F-measure of speech features for identifying UX problems for each test product.

	Coffee Machine			Website			Mobile App		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
High Loudness	0.38	0.21	0.27	0.67	0.23	0.34	0.55	0.18	0.27
High Pitch	0.42	0.22	0.29	0.54	0.19	0.28	0.46	0.15	0.22
High Speech Rate	0.38	0.05	0.09	0.53	0.06	0.10	0.58	0.05	0.09
Low Pitch	0	0	0	0.64	0.05	0.09	0.14	0.01	0.02
Low Loudness	0	0	0	0	0	0	0.21	0.02	0.03
Low Speech Rate	0	0	0	0	0	0	0	0	0

6.1 Key Takeaways

Our research quantified connections between verbalization and speech features and UX problems for older adults. It extended prior work, which uncovered subtle patterns indicating UX problems for young adults (e.g., [17, 24, 44]). Our findings show that the subtle patterns uncovered with young adults in prior work are largely applicable to older adults. First, the verbalization categories are similar for older and young adults. The observation category is the most indicative of UX problems. When users encounter a problem, their verbalizations tend to be of the Observation category than other categories.

Second, high speech rate, loudness, and pitch are indicative of UX problems. When users verbalize their thoughts in high loudness, pitch, or speech rate, they likely encounter a problem.

Third, while previous research suggested the role of sentiments in identifying UX problems with young adults [17, 24], our research quantified the correlation between the three types of sentiments and UX problems. Our findings show that while neutral sentiments are most common, negative sentiments are most indicative of problems.

Fourth, while prior studies highlighted most frequently verbalized words when users encountered problems [17, 24], our research revealed the most frequently verbalized words when users both encountered and did not encounter problems for better comparison. Specifically, users tend to verbalize negations, question words, and words with negative sentiments more often when they encountered problems; they tend to verbalize task-related nouns and verbs more often when they did not encounter problems.

Meanwhile, our findings also highlight differences. First, older adults' verbalizations have a higher proportion of Observation and a lower proportion of Reading than younger adults [17]. Second, older adults' verbalizations contain more negative sentiments and less positive sentiments than younger adults [24]. Third, while low speech rate had a high precision of indicating UX problems for young adults [17], our study did not find it for older adults.

6.2 Verbalization *Categories* and UX Problems

Our results showed that older adults' verbalizations included what they saw (i.e., Reading), what they did (i.e., Procedure), what they remarked (i.e., Observation), and their rationales for behaviors (i.e., Explanation), which were similar to the findings of prior work on young adults (e.g., [17, 24, 44]). Observation was the most frequently appeared category, followed by Reading, Procedure, Explanation, and Others. This trend was consistent with Fan et al's recent study [17] that was conducted with young adults.

However, the percentage of Reading and Procedure in our study (31.2%) was lower than that (56.3%) in their study; and the percentage of Observation (62.5%) was higher than that (37.6%) in their study. It seems to suggest that older adults tended to make remarks on test products relatively more often but verbalize what they saw and what they did on test products less often than young adults. There might be two potential reasons. First, older adults might have allocated more attention or cognitive resources toward the immediate tasks that they were working on due to natural motor and cognitive declines. Consequently, they had less cognitive resources to verbalize what they saw (i.e., Reading) or did (i.e., Procedure) when they were busy working on the tasks. Second, when older adults started to make remarks (i.e., Observation), they tended to slow down the tasks and thus had more resources to verbalize their thoughts.

Observation was the most indicative of UX problems in terms of precision, recall and F-measure. Although Explanation had lower F-measure and Recall than Reading or Procedure, it had higher precision. These trends were consistent with the findings of the think-aloud studies that Fan et al. conducted with young adults [17]. Compared to prior work [17, 24, 44], we had a new category "Others" to categorize segments that did not belong to the four categories. While there was a relatively small percentage of Others segments, Others had a relatively high precision for identifying UX problems.

6.3 Verbalization *Sentiments* and UX Problems

The most frequently appeared sentiment was neutral (61%). This suggested that participants' verbalizations, more often than not, did not exhibit positive or negative sentiments. What's more, the proportions of the three types of sentiments followed a similar trend for the three different products. Further, this number was also consistent with the result of the think-aloud studies that Hertzum et al. conducted with young adults [24]. Similarly, Hertzum et al. found that 58% to 61% of their participants verbalizations were neither positive nor negative [24]. While their participants' verbalizations had roughly equal proportions of negative and positive sentiments [24], our participants' verbalizations had more negative sentiments (31%) than positive ones (8%). Although it seemed to suggest that older adults tended to have more negative experiences than young adults, this conjecture needs further investigation because our study used different test products and was not conducted with the exact same experimental setups.

Our work is the first to quantify the correlation between the sentiments of think-aloud verbalizations and UX problems. Specifically,

negative sentiments were most indicative of UX problems, followed by neutral sentiments and then positive sentiments. Moreover, negative sentiments ($F - measure = 0.7$) were more indicative of UX problems than the Observation category ($F - measure = 0.63$). In addition, neutral and positive sentiments could also suggest UX problems. For example, sarcasms often carry a positive sentiment, but can actually indicate problems when interpreted in context.

6.4 *Speech Features* and UX Problems

High speech features were relatively more indicative of UX problems than low speech features. However, both high and low speech features had much lower recall and F-measure than the Observation category or Negative sentiments. One reason was that only a small percentage of segments had high or low speech features. Specifically, no more than 20% of the segments of any test product had any high or low speech feature. In contrast, there were on average 62.45% segments of the Observation category and 30.54% of the segments of negative sentiments.

High speech features (e.g., loudness, pitch, and speech rate) had a reasonably high precision (close to 0.5). This suggested that there was a close to 50% chance to find a problem if randomly checking a segment with high speech features.

Recent research with young adults found that low speech rate had a high precision of indicating UX problems [17]. However, participants in our study rarely verbalized their thoughts in low speech rate. As people's processing speed tend to slow down with aging [39], older adults may need to allocate relatively more attention and short-term memory toward the task when doing usability testing [11]. As Table 7 shows, there were few segments associated with low speech rate in our study. It seemed to suggest that with an overall slow speech rate, our participants rarely slowed down their speech even more though they occasionally sped up (i.e., high speech rate).

6.5 Top Frequently Verbalized Words

Our work extended previous research that only analyzed the top frequently verbalized words when participants encounter problems [17] by analyzing the top frequently verbalized words when participants both encountered and did not encounter problems. Our research showed that older adults tended to verbalize *negations*, *question words*, and *words with negative sentiments* more often when they encountered problems. This finding is consistent with the findings of the previous research with young adults [17]. These results together suggest that participants, young and old, tend to express their confusions with negations, questions, and negative words when encountering problems. Such verbalizations tended to be comments and remarks about their behaviors or the interfaces, which were more likely of the Observation category and were more likely associated with problems (Figure 4).

What's more, our research extended prior research and found that participants tended to verbalize *task-related nouns* and *verbs* more often when they did *not* encounter problems. This finding suggested that when participants did not encounter problems, they tended to verbalize what they were doing (e.g., task-related verbs and nouns). By definition, such verbalizations were of the Reading and Procedure categories, where were less likely associated with problems (Figure 4). In addition, participants verbalized negations (e.g. n't) and question words (e.g., how) to a much lesser extent when they did not encounter a problem. This happened when they restated the tasks (*"We don't want it to make coffee until 7:30"*) or when they read the instructions that contain question words (*"How to make drip coffee?"*).

6.6 Design Implications

Our research uncovers subtle verbalization and speech patterns indicating UX problems among older adults. First, these patterns, along with the ones uncovered with young adults (e.g., [17, 24, 44]), suggest that both what participants say and how they say it reveal UX problems. Specifically, UX practitioners should be alert when participants make observations (remarks), use negations, question words, and words with negative sentiments and when they raise or drop their voice's volume, pitch (tone), or speech rate.

Second, these subtle patterns could also inform the design of AIs (e.g., [16, 21]) to automatically capture potential UX problems encountered by a large number of participants, which are labor-intensive and time-consuming to find via traditional manual analysis methods [18, 33]. For example, companies and research labs could conduct and record think-aloud usability testing sessions for a product/prototype with participants via remote usability testing, which has been shown to be as effective as in-person lab studies [3, 7, 8, 42]. Moreover, global pandemics, such as COVID-19, might also accelerate the adoption of remote usability testing. Such a remote set-up allows for collecting more representative data from a large number of participants.

Once test sessions are conducted and recorded from participants, computational methods in speech processing, text mining and computer vision can be leveraged to extract the subtle patterns uncovered in this research. These patterns and UX problem labels of a portion of the test sessions can then be used to train an AI to detect problems in the remaining test sessions. However, it remains an open question of how best to build such AIs. For example, how to minimize the amount of data UX evaluators would need to label to train an AI? What types of information should the AI provide? In addition to detecting the occurrence of a UX problem, an AI might be more useful if it could explain the problem or express its confidence in its inference.

Last, UX evaluators are prone to the "evaluator effect" [25] and yet often do not have opportunities to analyze a test session with another evaluator in practice [20]. Our findings could inform the design of human-AI collaboration tools to mitigate the issue. For example, such AI-assisted tools could suggest the video segments of a test session signaling problems based on our findings (e.g., segments containing observations, negative sentiments or abnormal speech features). As a result, UX evaluators could better allocate their attention or gain a second perspective on their analysis from their AI "colleague." While recent research began to explore this area (e.g., [19]), it remains an open question of how best to design human-AI collaboration tools to facilitate UX test session analysis.

7 LIMITATIONS

Our study was conducted with a relatively small number of older adults. It is important to validate our findings with more older

adult participants. Further, our participants did not have physical or cognitive impairments. People with physical impairments might interact with computing systems differently than their counterparts. Consequently, their verbalizations might contain content related to their motor skills. Moreover, people with cognitive impairments might forget steps more easily than their counterparts, which might cause them to need to read instructions or repeat steps more often. Thus, it is important to further explore what and how people with motor or cognitive impairments verbalize their thoughts and uncover the telltale signs of problems that they encounter.

We intentionally included both physical and digital products as test products to evaluate the robustness of the subtle patterns. Moreover, we also included a digital website and a mobile app to cover a wide range of digital products. With this design, we were able to shed light on potential effects of products on the subtle patterns that indicate UX problems. However, as we only tested a limited number of products in our study, we could not conclude that subtle patterns are applicable to other products. Future research should investigate whether and to what extent products might affect the subtle patterns indicating UX problems.

Although we compared our findings with a similar previous study [17] to shed light on potential similarities and differences in subtle patterns for older adults and young adults, our participants worked on different test products as prior study (e.g.,[17]). Future work could conduct a controlled study with both older and young adults in the same setting to validate and extend our findings.

8 CONCLUSION

We have conducted concurrent think-aloud usability testing with older adults using physical and digital products. We have categorized participants' verbalizations, annotated sentiments, extracted speech features, and quantified how well these features were indicative of UX problems with precision, recall, and F-measure. We have also identified most frequently verbalized words when participants did and did not encounter problems to understand their verbalizations at a micro-level. In sum, our results showed that older adults' think-aloud verbalizations tended to be of the Observation category and included words with negative sentiments, negations, and question words when they encountered problems. In addition, when older adults verbalized their thoughts in high loudness, pitch, and speech rate, they likely encountered problems. In contrast, when older adults did not encounter problems, they tended to verbalize task-related nouns and verbs. What's more, these patterns were largely consistent across the three test products.

Our findings have confirmed and extended recent study that uncovered similar patterns with young adults [17]. Taken together, these findings suggest that subtle patterns in users' verbalizations and speech are telltale signs of UX problems regardless their ages. Future work could further validate and enrich these findings by examining other types of user data (e.g., facial expressions, eye-tracking data).

Our research offers three design implications for analyzing think-aloud usability test sessions. First, our research shows that subtle patterns in users' verbalization and speech features indicate UX problems they encounter. Thus, UX evaluators should pay attention to these patterns when analyzing think-aloud usability test sessions. Second, manually analyzing a large number of usability test sessions is labor-intensive and time-consuming. These subtle verbalization and speech patterns could be extracted from a portion of the usability test sessions to build AIs, which could help identify UX problems in the remaining test sessions efficiently. Lastly, these patterns could be utilized in the design of human-AI collaborative UX analysis tools (e.g., [19]). Instead of automatically detecting UX problems, such tools could play the role of an assistant by providing a different perspective to UX evaluators, who often do not have opportunities to analyze a test session with another UX evaluator in practice [18, 20, 33] but are prone to the "evaluator effect" [25].

REFERENCES

[1] Obead Alhadreti and Pam Mayhew. 2017. To intervene or not to intervene: an investigation of three think-aloud protocols in usability testing. *Journal of Usability Studies* 12, 3 (2017), 111–132.

[2] Obead Alhadreti and Pam Mayhew. 2018. Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[3] Morten Sieker Andreasen, Henrik Villemann Nielsen, Simon Ormholt Schrøder, and Jan Stage. 2007. What happened to remote usability testing? An empirical study of three methods. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1405–1414.

[4] Paul Boersma and David Weenink. [n.d.]. Praat: doing Phonetics by Computer. http://www.fon.hum.uva.nl/praat/.

[5] Ted Boren and Judith Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication* 43, 3 (2000), 261–278.

[6] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. "Why Would Anybody Do This?": Understanding Older Adults' Motivations and Challenges in Crowd Work. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2246–2257. https://doi.org/10.1145/2858036.2858198

[7] Anders Bruun, Peter Gull, Lene Hofmeister, and Jan Stage. 2009. Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1619–1628.

[8] Kapil Chalil Madathil and Joel S Greenstein. 2011. Synchronous remote usability testing: a new approach facilitated by virtual worlds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2225–2234.

[9] Jane Chung, Shomir Chaudhuri, Thai Le, Nai-Ching Chi, Hilaire J Thompson, and George Demiris. 2015. The use of think-aloud to evaluate a navigation structure for a multimedia health and wellness application for older adults and their caregivers. *Educational gerontology* 41, 12 (2015), 916–929.

[10] Lynne Cooke. 2010. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication* 53, 3 (2010), 202–215.

[11] Fergus IM Craik and Timothy A Salthouse. 2011. *The handbook of aging and cognition*. Psychology press.

[12] Joseph S Dumas, Joseph S Dumas, and Janice Redish. 1999. *A practical guide to usability testing*. Intellect books.

[13] F Ci Eisler. 1986. Cycle linguistics: Experiments in spontaneous speech.

[14] Sanne Elling, Leo Lentz, and Menno De Jong. 2012. Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE transactions on professional communication* 55, 3 (2012), 206–220.

[15] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data*. the MIT Press.

[16] Mingming Fan, Yue Li, and Khai N Truong. 2020. Automatic Detection of Usability Problem Encounters in Think-aloud Sessions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–24. https://doi.org/10.1145/3385732

[17] Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Truong. 2019. Concurrent think-aloud verbalizations and usability problems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5, Article 28 (2019), 35 pages. https://doi.org/10.1145/3325281

[18] Mingming Fan, Serina Shi, and Khai N Truong. 2020. Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *Journal of Usability Studies* 15, 2 (2020), 85–102.

[19] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N Truong. 2020. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26, 1 (2020), 343–352. https://doi.org/10.1109/TVCG.2019.2934797

[20] Asbjørn Følstad, Effie Law, and Kasper Hornbæk. 2012. Analysis in practical usability evaluation: a survey study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2127–2136.
[21] Julián Grigera, Alejandra Garrido, José Matías Rivero, and Gustavo Rossi. 2017. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies* 97 (2017), 129–148.
[22] Nielsen Norman Group. [n.d.]. Demonstrate Thinking Aloud by Showing Users a Video. https://www.nngroup.com/articles/thinking-aloud-demo-video/.
[23] Christian Herbst. 2019. A Python library for voice analysis. https://homepage.univie.ac.at/christian.herbst/python/namespacepraat_util.html.
[24] Morten Hertzum, Pia Borlund, and Kristina B Kristoffersen. 2015. What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction* 31, 9 (2015), 557–570.
[25] Morten Hertzum and Niels Ebbe Jacobsen. 2001. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction* 13, 4 (2001), 421–443.
[26] Masahiro Hori, Yasunori Kihara, and Takashi Kato. 2011. Investigation of indirect oral operation method for think aloud usability testing. In *International Conference on Human Centered Design*. Springer, 38–46.
[27] Man Huang, Derek Hansen, and Bo Xie. 2012. Older adults' online health information seeking behavior. In *Proceedings of the 2012 iConference*. 338–345.
[28] Petri Laukka, Clas Linnman, Fredrik Åhs, Anna Pissiota, Örjan Frans, Vanda Faria, Åsa Michelgård, Lieuwe Appel, Mats Fredrikson, and Tomas Furmark. 2008. In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior* 32, 4 (2008), 195.
[29] Clayton Lewis. 1982. Using the'thinking-aloud'method in cognitive interface design. *Research Report RC9265, IBM TJ Watson Research Center* (1982).
[30] Carolyn A Lin, Patricia J Neafsey, and Zoe Strickler. 2009. Usability testing by older adults of a computer-mediated health communication program. *Journal of Health Communication* 14, 2 (2009), 102–118.
[31] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).
[32] Tana M Luger, Thomas K Houston, and Jerry Suls. 2014. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *Journal of medical Internet research* 16, 1 (2014), e16.
[33] Sharon McDonald, Helen M Edwards, and Tingting Zhao. 2012. Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication* 55, 1 (2012), 2–19.
[34] Jakob Nielsen. 1994. *Usability engineering*. Elsevier.
[35] Erica Olmsted-Hawala and Jennifer Romano Bergstrom. 2012. Think-aloud protocols: Does age make a difference. *Proceedings of Society for Technical Communication (STC) Summit, Chicago, IL* (2012).
[36] Erica Olmsted-Hawala and Temika Holland. 2015. Age-Related Differences in a Usability Study Measuring Accuracy, Efficiency, and User Satisfaction in Using Smartphones for Census Enumeration: Fiction or Reality?. In *International Conference on Human Aspects of IT for the Aged Population*. Springer, 475–483.
[37] Alex Pentland. 2010. *Honest signals: how they shape our world*. MIT press.
[38] Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: how to plan, design and conduct effective tests*. John Wiley & Sons.
[39] Timothy Salthouse. 2000. *A theory of cognitive aging*. Elsevier.
[40] Siegfried Ludwig Sporer and Barbara Schwandt. 2006. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 20, 4 (2006), 421–446.
[41] Mr. Tappy. [n.d.]. Mr. Tappy - Mobile UX research made easy. https://www.mrtappy.com/.
[42] Tom Tullis, Stan Fleischman, Michelle McNulty, Carrie Cianchette, and Margaret Bergel. 2002. An empirical comparison of lab and remote usability testing of web sites. In *Usability Professionals Association Conference*.
[43] Tingting Zhao and Sharon McDonald. 2010. Keep talking: an analysis of participant utterances gathered using two concurrent think-aloud methods. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. 581–590.
[44] Tingting Zhao, Sharon McDonald, and Helen M Edwards. 2014. The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behaviour & Information Technology* 33, 2 (2014), 163–183.

9 APPENDICES

Table 9: The UX problems with usability heuristics violated and the corresponding problem descriptions and examples.

UX Problems (*Usability heuristics violated* [34])	Example problems and think-aloud verbalizations
Users could not receive appropriate feedback to know the system status in time (*Visibility of system status*)	**Food delivery app**: The price of the pizza after adding extra toppings was not obvious to users. For example, P6 was confused and verbalized, *"Ten dollars... so if I add it to the cart, does it say how many you can have?"*; **Coffee machine**: The indicator light on the power button misled the user. For example, when P7 saw the light was on, she thought she managed to program the coffee machine to make coffee at the set time in the future. However, she actually set the coffee machine to make coffee right away. *"There's a light...Oops, it's already starting to drip. Oh, no."*
The design did not speak users' language or failed to match users' mental model (*Match between system and the real world*)	**Food delivery app**: Users were confused about words used in the app. For example, P6 hesitated in choosing between "Carryout" and "Curbside pickup". She moved her finger back and forth between the two options and verbalized, *"OK..."*; **Coffee machine**: The annotation of the figures in the instructions did not match users' mental model. P10 was confused about the figure numbers, and verbalized, *"OK. Filter holder is... filter holder is 7... I thought 7 is the water."*
Users could not back out of a process or undo an action easily (*User control and freedom*)	**Food delivery app**: Users could not edit the items in the shopping cart. P5 verbalized, *"I don't know how to do the subtraction. I don't know how to take out of this thing."*; **Coffee machine**: When participants missed the digit they wanted, there was no easy way to undo it and they had to repeatedly press the button to loop back. P9 verbalized, *"Oops. How to go back? All the way around? I need 30. Up to 30. This thing takes forever."*
The design did not follow platform and industry conventions (*Consistency and standards*)	**Pet adoption website**: The adoption fee information was not displayed at a consistent position across all pages and thus was difficult to find. P5 verbalized, *"Pay fee. What is the pay fee? Why don't they just tell me the fee?"*; **Food delivery app**: Buttons for the same function were positioned inconsistently across the pages. P10 could not find the delete button to delete the item and verbalized, *"Alright. There gotta be a way to cancel that."*
The design failed to prevent problems from happening in the first place (*Error prevention*)	**Pet adoption website**: The website required certain formats for some input fields. However, it did not inform users about the formats or set appropriate constraints. P1 did not realize that he used the wrong format when entering location information and got confused:*"I did put it [address] in it [search bar]. What I suppose to do?"*; **Coffee machine**: The current time of the coffee machine's clock has to be reset before doing any operation when it is powered on. However, it is not highlighted in the instructions. P3 did not realize that he had not set the current time and was confused: *"I did exactly what it [the instructions] told me. Hold it for two seconds and before it stops flashing, set the time. It's not set."*
The design required users to memorize certain information (*Recognition rather than recall*)	**Pet adoption website**: The website did not save users' previous searches so they had to recall the options later on. P10 verbalized, *"Do we save it anywhere? Or does it automatically save?"*; **Food delivery app**: The total price of the items in the cart was not visible while ordering so users had to memorize the price and estimate whether it exceeds the budget. P8 verbalized, *"I got...I think somewhere around $60, so each pizza was $25. so I'm gonna put 4 toppings on this one."*
The design did not provide users with different ways to accomplish a task (*Flexibility and efficiency of use*)	**Pet adoption website**: The website did not offer users advanced search options, such as filtering pets' temperaments. P7 verbalized, *"We got a lot of cats here, but they are not telling me about the temperaments yet."*; **Food delivery app**: Users could not search for the items they wanted and had to go through the menus. P6 could not find the drink menu and verbalized, *"Where's that at? It's not there."*
The visual design failed to support users' primary goal effectively (*Aesthetic and minimalist design*)	**Pet adoption website**: The back button, served an important function but was not visually salient. P6 kept looking for the back button, *"Let's go back. Let's see where we are. Let's see."*; **Food delivery app**: The button size was too small for older users. P8 struggled to press on the button and verbalized, *"My big fingers don't work well."*
The design failed to provide users with effective error messages that could indicate problems and suggest solutions (*Help users recognize, diagnose, and recover from errors*)	**Pet adoption website**: The website did not provide the correct format when users made a mistake. P1 did not know how to make the address valid and tried again and again. *"Let's try search... Invalid... Go back... Try again... Come on."*; **Food delivery app**: When users entered invalid address, the error message did not provide enough information for users to recover the error. P10 was confused when the error message showed up: *"OK... [fall into silence]"*
Users could not get additional support to complete tasks from system documentation (*Help and documentation*)	**Coffee machine**: It was difficult to navigate the coffee machine's instructions. P7 verbalized, *"This is not an easy manual to navigate."*