SpeakFaster Observer: Long-Term Instrumentation of Eye-Gaze Typing for Measuring AAC Communication

Shanqing Cai
cais@google.com
Google LLC
Mountain View, California, USA

Subhashini Venugopalan
vsubhashini@google.com
Google LLC
Mountain View, California, USA

Katrin Tomanek
katrintomanek@google.com
Google LLC
Mountain View, California, USA

Shaun K. Kane
shaunkane@google.com
Google LLC
Mountain View, California, USA

Meredith Ringel Morris
merrie@google.com
Google LLC
Mountain View, California, USA

Richard J.N. Cave
rcave@google.com
Google LLC
Mountain View, California, USA

Robert L. MacDonald
bmacdonald@google.com
Google LLC
Mountain View, California, USA

Jon Campbell
joncamp@microsoft.com
Microsoft Research
USA

Blair Casey
blair@teamgleason.org
Team Gleason Foundation
New Orleans, Louisiana, USA

Emily Kornman
emily@teamgleason.org
Team Gleason Foundation
New Orleans, Louisiana, USA

Daniel Vance
daniel@teamgleason.org
Team Gleason Foundation
New Orleans, Louisiana, USA

Jay Beavers
dev@teamgleason.org
Team Gleason Foundation
New Orleans, Louisiana, USA

ABSTRACT

Accelerating communication for users with severe motor and speech impairments, in particular for eye-gaze-based augmentative and alternative communication (AAC) device users, is a longstanding area of research. However, observation of such users' communication over extended durations has been limited. This case study presents the real-world experience of developing and field-testing a tool for observing and curating the gaze typing-based communication of an eye-gaze AAC user with amyotrophic lateral sclerosis (ALS). With the intent to observe and develop technology to accelerate eye-gaze typed communication, we designed a tool and a protocol called the SpeakFaster Observer to measure everyday conversational text entry by the gaze-typing user, as well as several consenting conversation partners of the AAC user. We detail the design of the Observer software and data curation protocol, along with considerations for privacy protection. The deployment of the data protocol from November 2021 to April 2022 yielded a rich dataset of gaze-based AAC text entry from everyday life, consisting of 130+ hours of gaze keystrokes and 5,000+ curated speech utterances from the AAC user and the conversation partners. We present the key statistics of the data, including the speed (8.1±3.9 words per minute) and keystroke saving rate (-0.14±0.83) of gaze typing, patterns of utterance repetition and reuse, and the temporal dynamics of conversation turn-taking in gaze-based communication. We share our findings and also open source our data collection tools to further research in this domain.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; **Accessibility**.

KEYWORDS

gaze typing, augmentative and alternative communication, text corpus, conversation, context awareness, accessibility

ACM Reference Format:
Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Shaun K. Kane, Meredith Ringel Morris, Richard J.N. Cave, Robert L. MacDonald, Jon Campbell, Blair Casey, Emily Kornman, Daniel Vance, and Jay Beavers. 2023. SpeakFaster Observer: Long-Term Instrumentation of Eye-Gaze Typing for Measuring AAC Communication. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3544549.3573870

1 INTRODUCTION

In diseases that cause profound motor impairments such as amyotrophic lateral sclerosis (ALS) or cerebral palsy, if limb and speech motor control are lost then voluntary eye movement becomes the primary means of communication and interaction. In such cases, an on-screen keyboard driven by an eye tracker is often used to control an augmentative and alternative communication (AAC) tool. When coupled with speech generation (aka text-to-speech or TTS), eye gaze-based typing enables an alternative way of speaking to others. However, gaze typing is extremely slow, typically less than 10 words per minute (WPM) [19], an order of magnitude slower

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3573870

than fluent speech (at ≈ 190 WPM for English [33]). This presents a significant obstacle to effective communication and social well-being for gaze-typers [10]. In this study, we focus on observing and measuring the conversational communications of an eye-gaze AAC user over an extended duration with the intention of finding means to accelerate their communication.

Conducting research on the real usage of gaze-typers is difficult due to their health conditions (e.g., fatigue can impact willingness to learn new software), lack of mobility (e.g., hindering travelling to user study sites), slow communication (e.g., reducing the ability to give feedback on a new system), as well as potential invasions to their personal conversations (e.g., privacy concerns with logging system inputs). As a result of these practical challenges, few prior studies have measured or characterized gaze typing by motor-impaired users. Most studies employ able-bodied subjects as surrogates (e.g., [15, 18, 26, 30–32]) or rely on simulation or theoretical analysis of interface performance under optimal conditions (e.g., [13, 14]). Experiments on actual gaze-typers are usually short in duration and limited to lab settings (e.g., [20]). We are unaware of any studies reporting on long-term gaze typing use by disabled AAC users in natural settings.

Studying gaze typing by actual AAC users during everyday use is necessary to ensure that algorithms and interfaces tested in the lab and/or with non-disabled testers are properly tuned for real world use. For instance, on able-bodied users, research on typing in real world settings has yielded insights unobtainable from lab studies alone [5, 12, 21]. In particular, for AAC users, factors that may affect speed and accuracy such as fatigue or gradual drifts in eye-tracking calibration [27], practicing and reusing phrases [18, 32], or using adaptive word prediction and completion models [25, 29], may be missed in short lab studies. To limit the impact of these factors on a study's data, researchers may be forced to control for such factors ahead of time, e.g., by asking users to type non-repeating phrases (often generic text collected from non-AAC users, e.g., [17]) which again may not reflect daily use. Another potential benefit of studying everyday use of gaze typing by actual AAC users is the opportunity to collect corpora that can capture long-term trends and personalized patterns. The current lack of realistic AAC corpora has prompted researchers to assemble simulated corpora through crowdsourcing based on an imagined AAC scenario [28]. Further, if the goal of such corpora is to enhance communication for AAC users, it would help to include context from conversational partners [31]. Corpora from conversations involving real gaze-typing users would not only surpass imagined text in terms of representativeness, but may contain authentic clues about long-term trends, personalized vocabularies and grammar patterns, as well as contextual factors that impact communication.

Thus, there is a need for a reliable system of instrumenting eye-gaze typing during everyday use by AAC users over long periods (e.g., weeks or even months). Our interdisciplinary team comprised researchers and engineers with backgrounds in AI, HCI, and accessible technologies, speech language pathologists with clinical experience with AAC users, and members of a foundation working with people with ALS; this mix of perspectives resulted in our novel data collection software and protocol, with its strong emphasis on privacy and autonomy for the AAC user and their conversation partners. This case study presents the design and implementation of a system and protocol referred to as the *SpeakFaster Observer* for gathering, handling, curating, and analyzing AAC gaze-typing behavior in context. We describe the *Observer* tool and report our experience using it to curate and collect data from a consenting gaze-typing user who has ALS and the user's consenting conversation partners. The instrumentation of the participant's gaze-typing usage was conducted for a six-month period, yielding a rich dataset of keystrokes (>1,500 utterances) typed and embedded within the context of conversations (>5,000 utterances) with several partners.

Based on this dataset, we report several findings including metrics of gaze typing, patterns of phrase reuse, long-term trends, as well as aspects that are seldom studied in AAC communication and could differ significantly from typical speakers such as temporal dynamics and turn-taking [8, 23]. These results constitute the first extended study of AAC gaze-typing "in the wild" to our knowledge. We have open sourced the SpeakFaster Observer and hope that the tool helps in quantifying and benchmarking deployed gaze-typing solutions. Further, our ultimate goal is to accelerate communication for such users, so we hope that our study can help inform a baseline that can be used to evaluate and improve novel text-entry paradigms such as ones using neural network-based completions and expansions [1, 6, 22].

2 DESIGN OF SPEAKFASTER OBSERVER

The SpeakFaster Observer (or the *Observer* for short) is an application designed to gather conversational gaze-tying data (that the AAC user intends to use TTS to speak aloud) while providing the user full visibility and control over its state of operation. The Observer is implemented as a Windows® application that runs in the background, minimized as a system tray icon, and drives a user-provided USB LED to ensure a salient signal is visible to conversation participants during active data collection. The user can launch the application from a shortcut through eye-gaze clicking. Once launched, the application enters an inactive state, during which no data is logged and no signal is sampled from any sensors. This inactive state is indicated by the gray color of the icon in the system tray. Gaze clicking the icon toggles the Observer into the "active" state, as indicated by the red color of the icon. The application can be shut down anytime by the user through the "Exit" option in the context menu. Fig. 1 shows screenshots of the system tray icon under different states.

The active state of the Observer is divided into two distinct sub-states, *non-session* and *in-session*. The observer logs data only during the in-session state, in addition to a contextual lead time for audio data (described below). The transition from the non-session state to the in-session state is triggered by the user bringing a configurable *target application*, typically the user's primary application for speech generation through gaze-typing (e.g., Tobii Dynavox Communicator® and PRC Accent®), to the system foreground. The state reverts automatically from in-session to non-session if the target application has not been in the foreground for the last 5 minutes. This adds an extra layer of safety to prevent unwanted data collection if the user forgets to deactivate the Observer or when gaze-based conversation is paused for an extended period of time. The Observer gives the user full control over its state of operation,

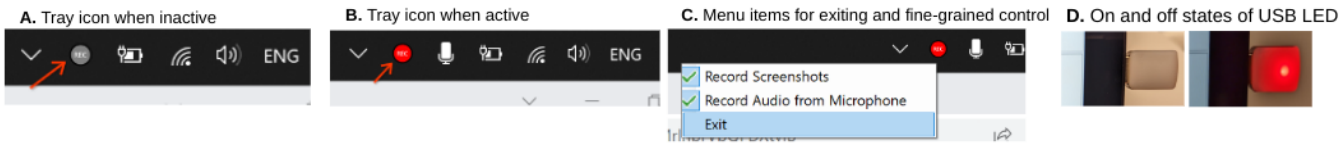

Figure 1: Panels showing the UI elements of the observer application. A, B, and C are screenshots and D includes photos of the USB LED under the off and on states.

including options to activate/deactivate at will, and exit at any time from any state (see the state diagram in Supplementary Fig. 1).

We now describe the data that the Observer gathers when it is active *and* in-session.

Keystroke logging: The Observer logs keystrokes only if the target application is in the foreground. While "*in-session*", the user may put another application in the foreground for a brief moment without causing the state to revert to *non-session*. No keystroke is logged during these moments despite the in-session state. This prevents keystroke logging unrelated to speech-based AAC, such as writing emails or entering passwords in a web browser. When the target application is in the foreground, the logged keys include all keystrokes entered by the user through gaze and ones issued by the on-screen keyboard on behalf of the user, which we refer to as *autokeys*. Autokeys, which are distinguished from human-issued keystrokes by their short spacing from the previous keystroke (<20 ms), occur when options for word completion and next-word predictions are selected and when the keyboard automatically performs an action that entails multiple keystrokes. For example, the `WordBackspace` feature in Tobii Dynavox keyboard [7] allows the user to conveniently delete the entire last word with a single gaze click (i.e., the sequence `Ctrl`, `Shift`, `Left`, and `Backspace`). Both content keys (alphanumeric, punctuation, and whitespace) and function keys (e.g., `Ctrl`, `Shift`) are logged.

Audio recording: When active, the Observer continuously pulls audio samples from the Windows system's primary microphone (16-bit mono, 16 KHz sampling rate). This audio is stored in a cyclic RAM buffer and not written to hard drive unless the Observer state is in-session, or if the audio is collected less than 5 minutes prior to the beginning of an in-session state. The rationale for logging audio 5 minutes prior to entering a session is to capture any utterances spoken by conversation partner(s) which the user's gaze typing may be a response to.

Screenshots: When the Observer is in-session and detects the target application in the foreground, it logs full-screen screenshots at a rate of 2 Hz. The collected screenshots are useful for the data curators (Sect. 3) to gain insight into user actions for text editing that are not available from the keystroke logs alone, such as mouse-based cursor navigation.

Storage and transfer. We refer to a continuous period of in-session state in the Observer as a *session*. The Observer creates a dedicated folder for each session, with the folder name as the start time of the session in UTC. Supplementary Table 1 summarizes the modalities, file formats, and estimated byte rate of data logged by the Observer during a session. The user can also delete logged sessions while they are saved on disk. When network access is available, the Observer periodically uploads the data from the local drive to a secure cloud storage bucket via an SSL connection. Once uploaded, the session folder is deleted permanently from the local device. This minimizes the risk of exhausting local disk space due to continuous collection; for instance, with the screenshots, data can be logged at a rate 24 MB/min. In our study, the user was also informed that uploaded data sessions could be deleted upon request by specifying the date and time of the collection (our user did not make any such requests).

3 PROCESSING AND CURATION OF OBSERVER DATA

The raw data collected by the Observer for each session undergoes automatic pre-processing, manual curation, and further automatic post-processing before it is made available to researchers (schematically shown in Fig. 2). During preprocessing, the session's audio files and screenshots are concatenated into a synchronized MP4 video file with *ffmpeg* [24]. The audio is transcribed using Google Cloud Speech-to-Text (automatic speech recognition or "ASR" hereafter) with the language set to U.S. English (our study participants' primary language). The utterance-level transcripts and logged keystrokes are written to a tab-separated values (TSV) file with timestamps consistent with the video file.

The manual curation of the data is then done via an open-source data annotation tool called ELAN[1] [3]. The ELAN software allows loading the MP4 video and the TSV file, and uses the timestamps to provide an interface where the screenshots, audio track, utterance-level speech transcripts, and keystrokes can all be viewed and edited in a time-aligned fashion. A trusted data curator (described in Sect. 4) then performs a six-step manual curation of each session. These steps include ensuring that speech transcripts from non-participating individuals are redacted, the speech utterances missed by ASR are added, ASR errors are detected and corrected or marked as background, the identity of the speaker for each utterance is labeled with pseudonyms, and that words in the participating individuals' utterances that carry sensitive and private information are redacted along with any corresponding keystrokes. The range of redacted contents not only included standard PII and sensitive topics such as health but also specific topics that the AAC user and his family requested to redact through discussion with the research team before the start of the data collection. The step-by-step instructions to the curators are listed in Supplementary Table 2.

Following the manual curation, a postprocessing script generates a single TSV file containing the timestamped keystrokes and speech transcripts with properly applied redaction. The script additionally generates a JSON file containing metadata regarding the session such as time zone, device information, and summary statistics of ASR errors. This pair of TSV and JSON files are purely textual, and constitute the *processed data* of the Observer session. The processed

[1] https://archive.mpi.nl/tla/elan

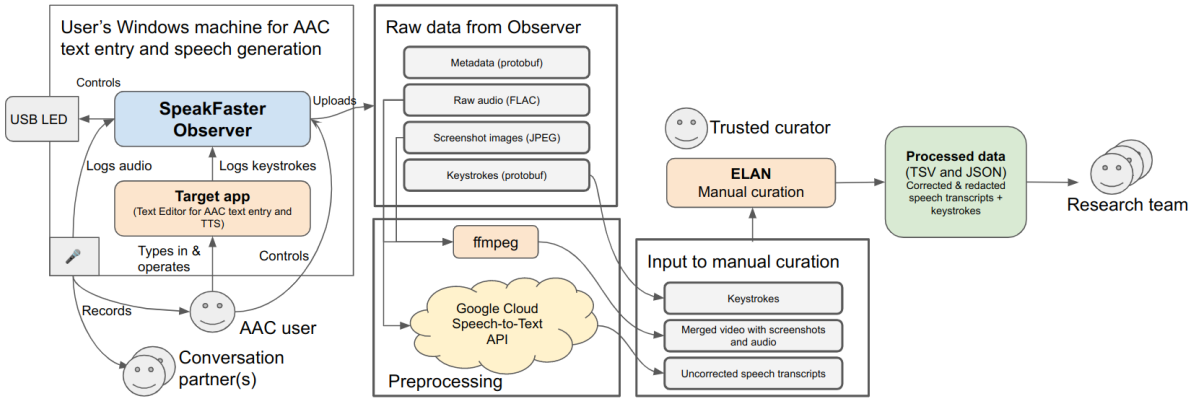

Figure 2: A schematic diagram illustrating the data flow in the SpeakFaster protocol. The SpeakFaster Observer collects raw data, which is automatically preprocessed, then manually curated via the ELAN tool, and is then postprocessed to yield purely textual TSV and JSON files for each session. The processed data is used in subsequent data analysis.

data from the recorded sessions forms the basis for subsequent research and analyses. This ensures that the raw audio-visual data from the Observer are exposed only to the trusted curators.

We have made the source code for the SpeakFaster Observer and the GUI tools and scripts for data preprocessing, curation, and postprocessing available in a public GitHub repository at https://github.com/TeamGleason/SpeakFaster.

4 AAC USER, CONVERSATION PARTNERS, AND PRIVACY CONSIDERATIONS

A single AAC user (also referred to as *User001*) participated in our case study. User001 is an adult male diagnosed with ALS and is a native U.S. English speaker. Five additional adults, including his spouse and four professional caregivers, were among the most frequently engaged conversation partners. All six provided written informed consent[2] following an introduction and question & answer session with the research team on the design of the data collection, processing, and curation procedures. A child of User001 was also a participating conversation partner, for which parental consent was obtained. Thus, one user and a total of six conversation partners participated in this study.

The team of trusted data curators consisted of three adults who had associate-level college education or above. Additional hiring criteria included trustworthiness, attention to details, and an interest in working with large amounts of text and audiovisual data related to everyday speech communication to help individuals with disabilities. They were interviewed by a trusted delegate of User001 during which they were informed about the purpose of the research program and introduced to the process of curation (Sect. 3). Given that the raw Observer data would contain daily conversations and some sensitive data, we conducted in-person and video introductions between the user's family and the curators before starting the data collection, to build trust and rapport. The curators' training included a tutorial on the data curation process, aided by hands-on practice with eight "practice" Observer sessions collected in two of the co-authors' own home environments, as well as audio recordings from the six conversation partners to familiarize their voice identities. The curators were paid $20 USD per hour. The curators reported that 2.5 to 4 hours were required to curate one hour of raw data (see Sect. 3), which is consistent with time ratios reported for manual speech transcription [4]. The data collection itself lasted six months, from November 2021 to April 2022.

4.1 User001's gaze-typing setup

User001 is an experienced gaze-typer who uses a Tobii Dynavox PCEye® Mini IS4 eye tracker (version 2.27.0) with Tobii Dynavox Windows Control® [7] to operate a Microsoft Surface® tablet. For speech generation, User001 uses Balabolka[3], a freeware text editor, and used the combo key `Ctrl+W` to activate TTS for the last-entered message consisting of one or more sentences. Balabolka supports multiple tabs, which User001 uses to store pre-composed utterances for different contexts. The user's speech output contains a mixture of novel phrases composed on the fly and phrases stored previously and re-spoken with or without modifications.

5 BASIC STATISTICS OF THE USER001 DATASET

We collected a total of 469 sessions from User001 and the six conversation partners over the six-month period. Seventy-four of the 469 sessions did not contain any speech utterances marked by the data curators as non-background and hence were removed from subsequent analyses. Statistics from the remaining 395 sessions are presented in Fig.3. The sessions are not uniformly distributed over time. There is a high volume of data in November, shortly after the commencement of collection (Fig.3A). The gap near the year-end holidays was related to the user preferring not to record due to other activities and the presence of many non-participating visitors. Fig. 3B shows the distribution of the sessions over the days of the week (x-axis) and 2-hour time bins over a day (y-axis). As expected, the sessions are distributed during the user's wakeful

[2]Consent form template is available at https://github.com/TeamGleason/SpeakFaster/blob/main/Legal/SpeakFaster_Data_Contribution_Agreement.md

[3]http://www.cross-plus-a.com/balabolka.htm

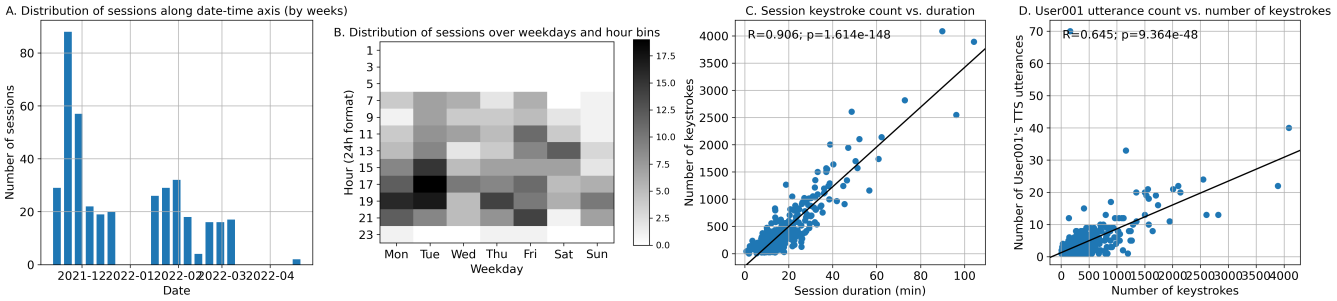

Figure 3: Temporal distribution and statistics from the observed sessions. A and B show distribution of the sessions over time of year and day respectively. C shows the correlation between the number of keystrokes with the session duration, and D the correlation between number of utterances vs. the keystrokes in the session; each dot in C and D represents an Observer session.

hours (7 AM to 11 PM). On weekdays, the sessions tend to gravitate toward the evening hours (7 PM - 9 PM), when the user converses with their family. We note that the observed sessions do not encompass all of User001's AAC communication in the time period, because the user had control over when the Observer is active and might likely reflect his preference for data collection, which was also influenced by the presence of the consenting conversation partners. In Fig.3C we see a significant correlation between the session duration (17.8±12.3 min, median=14.1 min) and the number of keystrokes (including autokeys and user-issued keystrokes) in the session (415±498, median=265, R=0.906, p=1.6e-148). Further, Panel D shows significant correlation between the number of keystrokes and the number of utterances spoken by User001's TTS (range: 1-70, 4.3±5.7, median=3, R=0.645, p=9.4e-48).

5.1 Speech utterances and conversation turn-taking

A total of 307 sessions contained at least one utterance spoken by User001 and one utterance from a consenting conversation partner. These sessions contained 5,504 utterances. Fig. 4A shows the fraction of utterances communicated by the different participants. User001's TTS output accounted for only 1,532 (27.8%) of these utterances. In addition to utterance count, the conversational imbalance was also seen in word counts (transcript character length / 5, [2]), where only 33.3% (22,217) of all words were produced by User001 while the rest (66.7%, 44,431) were by the conversation partners. Fig. 4B shows the fraction of sessions involving 1, 2, 3, or more conversation partners. About half of the sessions involved only one conversation partner. Fig. 4C shows that User001's average phrase length was greater than those of the oral conversation partners (14.5 words vs. 11.2 words, t-test: p=6e-8). This observation indicates that gaze-typing users do not necessarily regress to phrases shorter than spoken phrases uttered by non-disabled conversation partners.

Utterance repetition was a salient pattern of User001's TTS output and we analyze this in Fig. 4D. We define a repeating utterance as one with an identical transcript (normalizing whitespace, casing, and punctuation) to an utterance that the same speaker produced earlier in the same Observer session. Overall, 7.7% of User001's utterances were repetitions, significantly greater than the 2.5% repetition ratio for the conversation partners (χ^2 test: p=2e-41). Of User001's repeating utterances, a majority (75.4%) were repeating the immediately preceding utterance. Therefore the AAC user tended to repeat previous phrases 3x as much as non-AAC interlocutors. This may be related to our observation that conversation partners of the AAC user often temporarily disengaged and pursued other tasks while waiting for User001 to produce the next utterance. Therefore when User001's started speaking again via TTS, there was a frequent need to repeat in order to catch conversation partners' attention and to ensure that the new utterance was heard correctly. Balabolka, User00'1 TTS text editor, supported convenient repetition of the last spoken phrase.

5.2 Measuring the user's gaze-typed communication

To measure gaze typing, we examine the actual keystrokes issued by the user. We define an **utterance keystroke sequence (UKS, plural: UKSes)** as a consecutive sequence of keystrokes by which the gaze-typing user successfully enters an utterance and issues it as speech output via TTS. We require such a sequence to start from a content (non-function) keystroke and end with a key sequence that activates the TTS (Ctrl+W in User001's case). It may contain function keystrokes such as Backspace and WordBackspace. The text of the utterance can be reconstructed by taking into account the sequence of content and function keys. We analyze only those utterance keystroke sequences with non-empty reconstructed text. To exclude cases where the user abandoned or became distracted from the phrase entry, we exclude sequences where the time interval between any consecutive keystrokes exceeds 15 seconds.

We examined 856 UKSes from 320 Observer sessions which contained at least one UKS. The user issued 67,603 keystrokes among which 1.26% were redacted by the curators. Interestingly, these 856 UKSes accounted only for 33% of the total number of words in the utterances issued from the user's TTS indicating that only about a third of the user's spoken words were typed on the fly during the sessions and the rest were re-activations of previously stored phrases in Balabolka. Fig. 5A shows that the inter-keystroke intervals (IKIs) of the UKSes follow a bimodal distribution, with a sharp peak of extremely small values (<20 ms) corresponding to autokeys, along with a wider peak with values greater than 500 ms (User001's gaze-typing dwell time) and a long tail reaching 4 seconds and higher. Autokeys account for 48.3% of all the keystrokes in

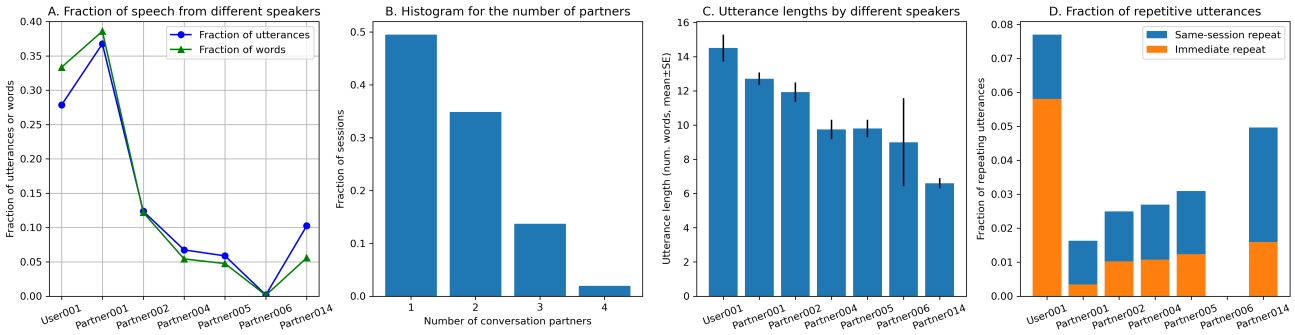

Figure 4: Statistics of speech utterances from User001 (via TTS) and the consenting conversation partners. A: How the speech data is distributed among User001 and the six conversation partners in terms of utterance count and word count. B: The distribution of the sessions in terms of the number of conversation partners present. C: Average utterance lengths of the user and partners. D: Utterance repetition by the user and partners.

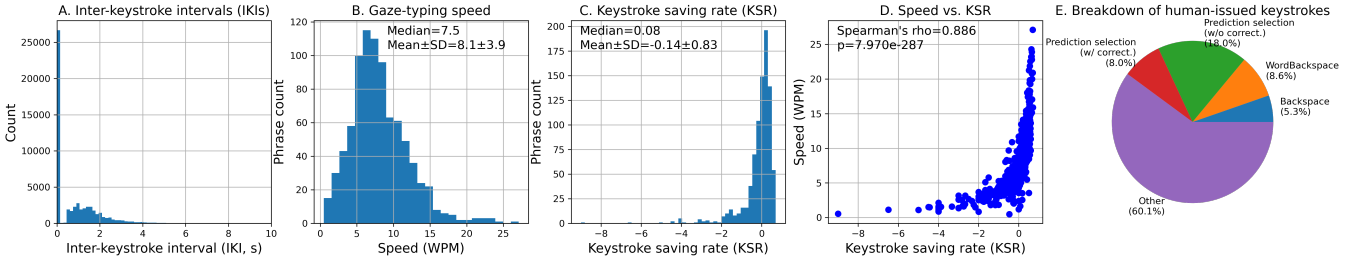

Figure 5: Measuring User001's gaze typing. The distribution of the inter-keystroke intervals (A) shows a bimodal pattern for autokeys and human-issued keys. Distribution of text-entry speed (B) and KSR (C) of the user's utterances are shown. A positive correlation (D) is observed between the text-entry speed and KSR. (E) distributes human-issued keystrokes by type.

the UKSes. Fig. 5B shows the distribution of the speed in words per minute (WPM) of all utterance keystroke sequences. We see that the dispersion is over a wide range from near zero to occasional values over 20 WPM, and the average text-entry rate is 8.1±3.9 WPM.

To measure the economy of keystrokes during utterance entry, we use the metric keystroke saving rate (KSR), defined as $KSR = \left(1 - \frac{N_H}{N}\right)$ where N is the number of characters in the utterance and N_H is the number of human-issued keystrokes (i.e., non-autokeys) including the function keys required to trigger the TTS output such as `Ctrl+W`. KSR has a theoretical ceiling of 1 corresponding to the ideal limiting case of outputting a TTS utterance with zero keystrokes. A negative KSR value indicates that number of human-issued keystrokes exceeds the character count of the utterance. The KSRs from the 856 UKSes follow a distribution skewed towards large negative values (Fig. 5C). Although the median KSR is 0.08, the mean is negative (-0.14±0.83). This skew is due to a number of utterances containing significant edits, i.e., many `Backspaces`- and `WordBackspace`-based corrections. Overall, there is a significant correlation between KSR and text-entry speed (Spearman's ρ=0.886, $p \approx 0$), indicating a strong association between greater saving of keystrokes and faster text entry at the utterance level (Fig. 5D).

Over the six-month period, we did not observe significant month-to-month variations in the gaze-typing speed (p>0.2) or KSR (p>0.8), indicating a stable performance at the large timescale. However, there were trends of significant variations in the speed (p=0.011) and KSR (p=0.063) with the time of the day. The afternoon hours (14-16) saw the highest speed and KSR, while the morning hours (8-10) saw the lowest speed and KSR (Supplementary Fig. 2)

The user's typo corrections was performed primarily through `Backspace` and `WordBackspaces`, which accounted for 13.9% (Fig. 5E) of all human-issued keystrokes. This ratio was considerably higher than the ratio of `Backspaces` reported for mobile text entry (8.9%, [5]). Fig. 5E also shows that approximately 26% of the keystrokes issued by the user were for selecting word predictions (including word completion and next-word prediction). Among these selected options, a large fraction (8% out of 26%) involve erasing incorrectly-entered prefix letters (e.g., typing `h` and `l`, followed by selecting the word-completion option `half`, involves a `Backspace` autokey that first erases the preceding `l`).

6 DISCUSSION

We designed SpeakFaster Observer, and our research protocol, to protect the privacy of the gaze-typing user and their conversation partners. First, the software granted the user full control over when to collect data; we saw that the user valued this control by choosing to focus their data collection on a subset of their daily interactions

(e.g., interactions in their own home during the evening). It communicates the data collection status clearly to the user as well as the conversation partners present through both the UI (system tray icon color) and hardware-based (LED) visual cues. A two-tier organization protected the raw data, wherein only the trusted curators had access to the raw keystrokes and audio-visual data. The curators removed and redacted unrelated and sensitive information, producing clean text-only data that was shared with researchers for subsequent analyses. While this approach has clear benefits for protecting the privacy of the AAC user and their conversation partners, it also has drawbacks – the approach is time-consuming (requiring 2.5-4 hours of manual data curation per 1 hour of raw data), expensive (requiring $65 USD per hour of raw data to pay the data curators), incurs cognitive cost (the AAC user must decide when to turn the data collection on and off), and reduces the completeness and generalizability of the data collected.

One of the goals of the study was to test the feasibility of collecting a text corpus for AAC through instrumenting gaze typing. Through keystroke logging, we obtained a corpus of 6,000+ words over a period of six months. In comparison, approximately three times as many words were available via the ASR transcripts based on the audio recordings in the same sessions. This shows that our user frequently reused text entered previously (i.e., not a part of the Observer sessions), for which the text data was unavailable through keystroke logging. Keystroke logging has the additional limitation of not capturing non-keystroke UI events related to text entry, including moving the cursor or selecting text (performed by gaze users through software such as Tobii Windows Control [7]). Data curators could use screenshots captured by the Observer to manually annotate such UI events, but we decided against giving curators yet another task on top of their already labor-intensive workflow. Instead we used the audio recordings and the downstream ASR transcripts as the ground truth text corpus. These gaps in data collection are caused by the design of AAC applications, and would likely appear when collecting data from other users. Therefore future corpus collection for gaze-typing users should rely on a more direct instrumentation path (e.g., via the speech-generating application's own API) to achieve the highest possible coverage and efficiency of corpus collection. However, such APIs are not always available, and tailoring software to each system's API reduces the generality and reusability of such systems. A more generalize alternative is to apply ASR to the user's TTS output through a loop-back audio channel.

While creating a publicly available corpus of real-world, longitudinal AAC data would greatly benefit the research community, there remain several obstacles to achieving this goal, including cost concerns related to data cleaning on such a large scale and privacy concerns. Since our corpus is only from a single user, we have chosen not to share it publicly due to privacy concerns. Even if data were able to be collected and combined from a larger set of users (thereby increasing k-anonymity), serious privacy challenges would remain – paying data curators at a large scale may be economically infeasible, and automated techniques to scrub sensitive information at that scale would be imperfect, thus leaving open the potential for re-identification of AAC users and/or their conversation partners, as well as potentially leaving sensitive information in the data set. We hope that our case study reflections and accompanying Observer software are a first step toward helping the community consider the trade-offs and feasibility of creating a safe and ethical public corpus of authentic AAC data.

During conversations, the gaze-typing user's speech output accounted for less than a third of the total amount of speech from all participating interlocutors. Further, just a third of those utterances were composed on-the-fly via gaze typing; the remainder were reactivations of previously-entered phrases. This high ratio of reuse hints strongly at the benefit of context-dependent phrase retrieval (e.g., [13]). For phrases entered on-the-fly, the observed 8.1-WPM average text-entry rate remained stable over the six-month of period and is consistent with previously reported ranges of 5-10 WPM for gaze-based AAC users [19]. However, considerable speed variability existed from utterance to utterance. While the rate occasionally reached 20 WPM (1.4% of utterances, Fig. 5), many utterances suffered from rates less than 2 WPM (2.9% of utterances), which is attributable to the large number of keystrokes required to fix gaze typos (Fig. 5E) reflected as negative KSR values (-0.14 on average, much lower than the range of 0.15-0.2 for tap-based mobile typing by able-bodied users [16]). The fact that User001 painstakingly fixed typos while maintaining utterance lengths slightly exceeding those of able-bodied partners is consistent with previous observations that many AAC users strive to preserve the quality and personality of their language at the cost of greater effort and lower speed [10]. Our data shows a strong positive correlation between KSR and text-entry rate, indicating that efforts to improve the accuracy of eye tracking and exploring novel approaches of reducing the number of keypresses required (e.g., through context awareness and abbreviation paradigms [6, 22]) are important directions in future gaze-typing AAC research.

7 CONCLUSION AND FUTURE DIRECTIONS

In this study, we presented the SpeakFaster Observer, a software application and data handling protocol to gather conversational communication data of an AAC user. Our experience demonstrates that this can be unobtrusively deployed to an AAC user without affecting their habitual communication. Through thoughtful data curation, the Observer yielded keypress and conversation data that allowed us to gain useful insight into the speed, efficiency, and temporal patterns of gaze typing in everyday usage of a gaze-based AAC user with ALS.

The contextualized text corpus collected from User001 opens the door to future case studies on the adaptation of n-gram and neural language models [11], abbreviation expansion [6, 22], and context-dependent phrase retrieval [13] based on real user data. Although the Observer is designed as an observational tool, it is also an attempt at testing the feasibility of a multi-modal context-aware AAC text-entry application from technical and privacy perspectives. Our current protocol analyzed the speech in an offline fashion; future studies can explore online ASR of conversation partners' speech. Building on the Observer, future studies can also explore other modalities of contextual data, such as images from cameras of the eye-gaze computer and geolocation, which may contain useful contextual signals for improving text prediction [9]. A context-aware AAC application that omits humans in the loop (the data curators) will have the benefits of greater automation, lower cost, and less

persistence and exposure of data to other humans, but will entail the challenges of greater error rate (e.g., in speech transcription). The tradeoffs between privacy, predictive power, and cost in AAC applications is a nascent field and one which our case study can shed some light on.

ACKNOWLEDGMENTS

We thank the data curators, Chris Flood, Giovanni D'Amico, and Colby Workman, for their dedicated work. Sean Holmes provided technical assistance to the data curators. Julie Cattiau assisted with preparing the training materials for data curation.

REFERENCES

[1] Jiban Adhikary, Jamie Berger, and Keith Vertanen. 2021. Accelerating Text Communication via Abbreviated Sentence Input. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6574–6588.

[2] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2009. Analysis of text entry performance metrics. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*. IEEE, 100–105.

[3] Eric Auer, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, Stefano Masnieri, Daniel Schneider, and Sebastian Tschöpel. 2010. ELAN as flexible annotation framework for sound and image processing detectors. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*. European Language Resources Association (ELRA), 890–893.

[4] Thierry Bazillon, Yannick Esteve, and Daniel Luzzati. 2008. Manual vs assisted transcription of prepared and spontaneous speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

[5] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A mobile keyboard application for studying free typing behaviour in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[6] Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Ajit Narayanan, Meredith R Morris, and Michael P Brenner. 2022. Context-Aware Abbreviation Expansion Using Large Language Models. *arXiv preprint arXiv:2205.03767* (2022).

[7] Tobii Dynavox. 2017. *Tobii Dynavox Windows Control User's Manual*. http://tdvox.web-downloads.s3.amazonaws.com/Windows%20Control%202/Documents/TobiiDynavox_WindowsControl2_UserManual_v1-0-1_en-US_WEB.pdf

[8] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634.

[9] Shaun K Kane and Meredith Ringel Morris. 2017. Let's Talk about X: Combining image recognition and eye gaze to support conversation for people with ALS. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 129–134.

[10] Shaun K Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. " At times avuncular and cantankerous, with the reflexes of a mongoose" Understanding Self-Expression through Augmentative and Alternative Communication Devices. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1166–1179.

[11] Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 2461–2469.

[12] Andreas Komninos, Mark Dunlop, Kyriakos Katsaris, and John Garofalakis. 2018. A glimpse of mobile text entry errors and corrective behaviour in the wild. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 221–228.

[13] Per Ola Kristensson, James Lilley, Rolf Black, and Annalu Waller. 2020. A design engineering approach for quantitatively exploring context-aware sentence retrieval for nonspeaking individuals with motor disabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[14] Per Ola Kristensson and Keith Vertanen. 2012. The potential of dwell-free eye-typing for fast assistive gaze communication. In *Proceedings of the symposium on eye tracking research and applications*. 241–244.

[15] Andrew Kurauchi, Wenxin Feng, Ajjen Joshi, Carlos Morimoto, and Margrit Betke. 2016. EyeSwipe: Dwell-free text entry using gaze paths. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 1952–1956.

[16] Tianshi Li, Philip Quinn, and Shumin Zhai. 2022. C-PAK: Correcting and Completing Variable-length Prefix-based Abbreviated Keystrokes. *ACM Transactions on Computer-Human Interaction* (2022).

[17] I Scott MacKenzie and R William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI'03 extended abstracts on Human factors in computing systems*. 754–755.

[18] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast gaze typing with an adjustable dwell time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 357–360.

[19] Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications*. 15–22.

[20] Martez E Mott, Shane Williams, Jacob O Wobbrock, and Meredith Ringel Morris. 2017. Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2558–2570.

[21] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 679–688.

[22] Junxiao Shen, Boyin Yang, John J Dudley, and Per Ola Kristensson. 2022. Kwickchat: A multi-turn dialogue system for aac using context-aware sentence generation by bag-of-keywords. In *27th International Conference on Intelligent User Interfaces*. 853–867.

[23] Kiley Sobel, Alexander Fiannaca, Jon Campbell, Harish Kulkarni, Ann Paradiso, Ed Cutrell, and Meredith Ringel Morris. 2017. Exploring the design space of AAC awareness displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2890–2903.

[24] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.

[25] Keith Trnka. 2008. Adaptive language modeling for word prediction. In *Proceedings of the ACL-08: HLT Student Research Workshop*. 61–66.

[26] Outi Tuisku, Päivi Majaranta, Poika Isokoski, and Kari-Jouko Räihä. 2008. Now Dasher! Dash away! Longitudinal study of fast text entry by eye gaze. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 19–26.

[27] Mindaugas Vasiljevas, Thomas Gedminas, A Ševčenko, M Jančiukas, Thomas Blažauskas, and Robertas Damaševičius. 2016. Modelling eye fatigue in gaze spelling task. In *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 95–102.

[28] Keith Vertanen and Per Ola Kristensson. 2011. The imagination of crowds: conversational AAC language modeling using crowdsourcing and large data sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 700–711.

[29] Tonio Wandmacher, Jean-Yves Antoine, Franck Poirier, and Jean-Paul Départe. 2008. Sibylle, an assistive communication system adapting to the context and its user. *ACM Transactions on Accessible Computing (TACCESS)* 1, 1 (2008), 1–30.

[30] David J Ward, Alan F Blackwell, and David JC MacKay. 2000. Dasher—a data entry interface using continuous gestures and language models. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*. 129–137.

[31] Bruce Wisenburn and D Jeffery Higginbotham. 2008. An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: Objective results. *Augmentative and alternative communication* 24, 2 (2008), 100–109.

[32] Jacob O Wobbrock, James Rubinstein, Michael W Sawyer, and Andrew T Duchowski. 2008. Longitudinal evaluation of discrete consecutive gaze gestures for text entry. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 11–18.

[33] K Yorkston, David Beukelman, and R Tice. 1996. Sentence intelligibility test. *Lincoln, NE: Tice Technologies* (1996).