# "I Want to Publicize My Stutter": Community-led Collection and Curation of Chinese Stuttered Speech Data

QISHENG LI, AImpower.org, USA

SHAOMEI WU, AImpower.org, USA

This paper documents the process undertaken by *StammerTalk*, a grassroots community of Chinese-speaking people who stutter, to autonomously collect and curate stuttered speech data for more inclusive speech AI models. While people with disabilities are often excluded or treated merely as the *subjects* of AI data collection, our work introduces a new model for disability data collection in which the disability community exerts agency and control over their personal data and data-driven experiences. Our ethnographic data show that community-led data collection not only produces data needed to represent the community in AI systems, but also empowers the community and its members, by embracing - rather than concealing - stuttering and stutterer identity, and strengthening the social bonds of the community. Recognizing the lack of adequate socio-technical infrastructure for community-led, grassroots data collection, we discuss practical challenges, as well as the strategies and factors for communities to succeed in similar endeavors.

CCS Concepts: • **Human-centered computing** → **Empirical studies in accessibility**; *Accessibility theory, concepts and paradigms*; • **Computing methodologies** → **Language resources**; **Speech recognition**.

Additional Key Words and Phrases: AI FATE, datasets, data practice, community data model, representation, speech technology, disability, accessibility, stuttering, stuttered speech

**ACM Reference Format:**

Qisheng Li and Shaomei Wu. 2024. "I Want to Publicize My Stutter": Community-led Collection and Curation of Chinese Stuttered Speech Data. In . ACM, New York, NY, USA, 25 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

While the rapid progress of Artificial Intelligence (AI) in vision, language, and creative tasks promises innovative and powerful assistive technologies benefiting people with disabilities (PWD) in the future, the current landscape of AI technologies presents numerous challenges and threats to the lives of PWD today. Such challenges and threats include ableist microaggressions [14, 19], degraded quality of services [24, 47], additional accessibility barriers [24], and censorship of disability content [3, 19]. In general, the needs and requirements of PWD have not been prioritized in AI technologies, as they were developed without the active involvement of the disability community [14, 41], overlooking a crucial principle of the disability rights movement – "*Nothing About Us Without Us*" [9].

As popular AI technologies - such as large language models (LLMs) and generative AI (GAI) - often relies on big data, the inadequate and often biased representations of PWD in AI datasets has been identified as a fundamental issue that contributes to biases and discrimination towards PWD observed in various AI models [14, 19, 29, 41, 47]. Collecting data from and about PWD has been a challenge for the AI community: not only limited in size and socioeconomic status, PWD are also often excluded from data collection due to physical and digital accessibility barriers [29]. Some recent efforts have been made to include people with disabilities in AI data [17, 21, 28, 29, 37]. However, sponsored by tech



Manuscript submitted to ACM

companies [28] or academic institutions [17, 21, 29, 37], current efforts have primarily been orchestrated by external "experts" rather than by the disability community itself, and often treated people with disabilities as *data subjects* rather than the *owner* and *controller* of the collected data [42]. Even when the data are collected with a participatory approach (e.g. [17, 37], participants usually have rather limited decision power about the data collection and usage, often taking a passive role of being informed and consulted [12]. Essentially, the expert-led data model deprives people with disabilities with their agency and control over their personal data, making it difficult to engage and incentivize the disability community to participate in AI data collection [29].

The emerging practice of community-driven, grassroots data collection presents opportunities for marginalized communities to exert agency and control over their personal data and data-driven experiences [1]. While individuals might lack the power to influence large AI models, collectively, the disability community is both self-motivated and capable to co-create authentic and adequate datasets about themselves to undo algorithmic biases and harms. In this paper, we present a case study for the community-driven, grassroots AI data collection initiative led by StammerTalk, an online community for Chinese-speaking people who stutter (PWS). Frustrated by the poor performance of automatic speech recognition (ASR) systems for stuttered speech [24], the StammerTalk community self-organized to create and curate the first and largest Chinese stuttered speech corpus to improve their experience with speech AI technologies. As a third-party partner to StammerTalk, we closely followed the community's progress from the inception of the initiative, collecting rich ethnographic data through our notes and observations. We also conducted interviews and surveys with community members to understand the process, benefits, and challenges of their grassroots efforts to collect disability-related data for fair and inclusive AI models. This study provides an in-depth look at community-led data collection processes and their implications for AI development.

Our study shows that the community-led AI data practice not only produce the much needed data to authentically represent the disability community in AI systems, but also benefits the community and its members beyond the tangible technical outcomes. Contrary to what was observed in expert-led data collections [29], StammerTalk members who participated in the speech data collection were driven by intrinsic goals - such as the making meaningful contribution to the community and bonding with other people who stutter, rather than monetary compensation. Community participants also found the data collection process pleasant and satisfying, enjoying the unique experience to talk about stuttering and their experience as a person who stutters in a safe and empathetic space. Beyond the positive experience during data collection, community participants also reported gaining valuable communication skills and deeper insights on stuttering, finding a sense of empowerment and stronger communal bonds beyond the data collection sessions.

Our study also uncovers challenges the StammerTalk community faced, as a result of limited resources and lack of adequate socio-technical infrastructure for grassroots data initiatives by marginalized communities. Besides the time and energy required for community members to design, execute, and quality control the data collection process, they also needed to navigate regional and cross-border data regulations - which often come with complex geo-political implications - when working with geographically distributed community members and partners across the world.

Taken together, our work illustrates the benefits and challenges of grassroots community AI data initiatives, and calls for the attention and investment from industry practitioners, academic researchers, and policymakers to develop socio-technical solutions that support broader adoption of such data practice, as it not only produces critical data for fair and inclusive AI models for PWD, but also serves data justice for the disability community.

## 2 RELATED WORK

To contextualize our work within the existing literature, we cover prior work on AI biases and discrimination against people with disabilities, with a focus on stuttering and speech AI. We then review existing efforts to include people with disabilities into AI datasets, discussing their limitations and challenges. Finally, we offer an overview of the emerging research and practice of alternative data models, under the framework of data justice.

### 2.1 AI Fairness Challenges for PWD

As race and gender based biases and discrimination in AI models become salient [7, 8, 31, 46], researchers and disability advocates have also identified AI fairness issues regarding people with disabilities.

One prominent concern is the performance disparities observed in AI models when interacting with people with disabilities. For instance, trained over photos taken and uploaded by sighted people [13], computer vision models frequently fail to accurately classify, recognize, and describ photos taken by people with visual impairments [17, 47]. Similarly, popular Automatic Speech Recognition (ASR) models were shown to perform drastically worse when transcribing the speech from Deaf and Hard-of-Hearing (DHH) people [15].

Beyond performance disparities, AI systems can also impact PWD by reinforcing existing social stigma and facilitating systematic marginalization. A recent study of LLMs from the perspectives of people with disabilities revealed that the conversational responses from the model "*mirrored subtle yet harmful stereotypes*" about PWD [14]. YouTubers with disabilities have reported constraints imposed by content distribution algorithms, limiting their reach to a wider, general audience [10]. More overtly, Hutchinson *et al.* found that content moderation algorithms systematically over-predicted disability-related text as toxic [19].

In the realm of stuttering and AI, the challenges are particularly pronounced in speech technologies. Despite the popularity and benefits of ASR-powered speech interfaces, recent research has shown that ASR systems struggle to understand stuttered speech, exhibiting a three to four times higher word error rate (WER) compared to non-stuttered speech [24]. In practice, ASR systems are more prone to misinterpreting the speech of PWS, cutting them off prematurely, and failing to respond correctly [5]. The inability of ASR systems to process stuttered speech could make it extra difficult for PWS to interact with smart speakers, automatic phone menus, in-car navigation systems, creating structural barriers and emotional distress that further marginalize them in our society.

### 2.2 Creating Representative AI Datasets for PWD

Researchers have converged on the idea that the lack of representative data from and about people with disabilities in AI training and testing poses a bottleneck for developing fair and inclusive AI models [14–16, 19, 41, 47]. In response, AI researchers and companies have undertaken numerous efforts to create disability-specific AI datasets.

One approach involves adapting data about PWD from other domain applications for AI purposes. For example, images uploaded to VizWiz, an application for visually impaired users to crowdsource answers to visual questions [4], were annotated and used to train computer vision models to better recognize photographs by people with visual impairments [17]. In the context of stuttering and ASR, the FluencyBank dataset [32], primarily collected to educate and train speech language pathologists, is frequently used for benchmarking and tuning ASR models for stuttered speech [25]. Recordings of podcasts by people who stutter were collected and repurposed to train ASR models to detect stuttering events in speech [25]. While this approach could be cost effective, it also presents challenges. First, depending on the original use case, the datasets may not easily match the needs of today's AI models regarding size, format, and

labels [17, 25]. Second, although collected with explicit consent for the original use case, it is unclear whether the original participant agreements extend to other use cases or broader data sharing.

Another popular approach is to generate synthetic disability data by simulating disability conditions with general population data. For instance, Wu *et. al.* injected writing errors frequently occurred in writings of Facebook users with dyslexia into millions of randomly sampled posts on Facebook [44] to train a spell and grammar checking model for users with dyslexia. LibriStutter [23], a popular stuttered speech dataset, was created by injecting synthetic stutters (repetitions, prolongations, interjections) into recordings of fluent speech. Sharing the general issues with disability simulations [22, 34], this approach is limited to capture the authenticity and diversity of the disability community to fairly present PWD in AI data.

Recent work explores the approach of collecting data directly from the disability community for AI purposes. Theodorou *et al.* designed a mobile App for users with visual impairments to take photos of objects to train a Teachable Object Recogniser [38]. Park *et al.* experimented with an online portal for participants with disabilities to upload data such as photos, speech, and videos, contributing to AI datasets [29]. Through Project Euphonia, a web interface that allows people with speech impediments to record and upload their speech samples, Google has collected over 1400 hours of atypical speech data to improve their ASR models [28]. while promising, this approach faces challenges in providing resources and assistance needed during data collection, quality control of the collected data, motivating and retaining data contributors, and addressing heightened privacy concerns over sensitive personal data [6, 29].

Fundamentally, current approaches follow an "expert-led" model, where experts like AI researchers and companies (*data controllers*) dictate what and how data about the disability community is collected, used, and shared. The community is often considered merely as *data subjects*, with little agency or legal rights over their personal data once in the hands of large institutions and corporations. This power imbalance determines that the data collection effort would inevitably become a transaction through which the experts paying the disability community for their data, rather than a meaningful partnership. As a result, the data collected often fail to represent the disability community fairly and adequately, due to the lack of trust, incentives, and intellectual inputs from the community [42].

### 2.3 Data Justice and Alternative Data Models

To transform "*existing power asymmetries and inequitable or discriminatory social structure*" regarding personal data [26], legal and policy scholars have introduced the concept of data justice, framed by six pillars: power, equity, access, identity, participation, and knowledge [26].

Under the data justice framework, new legal (e.g. European Union General Data Protection Regulation) and technological tools (e.g. Data Transfer Project[1]) have been developed for data subjects to control and manage their data. However, operating at the individual level, these tools often require extensive legal knowledge and technological resources that people with disability could rarely afford [42].

Recently, alternative data models, such as data trusts [36], data foundations [36], data cooperatives [36], data commons [33], and data sovereignty [40], have emerged to facilitate collaborative personal data stewardship within communities. While designed to provide data subjects with more agency over the collection and use of their personal data, these data models come with practical challenges. Most of them require significant operational, legal, and technical resources to deploy. Some, like data trusts, remain largely theoretical [42].

[1]https://dtinit.org/

Some technical solutions have been created to explore these data models today. For example, Driver's Seat[2] is a mobile app that enables rideshare and delivery drivers to share their driving data in a data cooperative to optimize work time and earnings. However, such applications are often domain-specific, with well-defined user goals and values. It remains unclear whether similar applications can be designed to collect and manage data for training foundational AI models, a use case that is more open-ended and without tangible, immediate benefits for individual users.

Given StammerTalk's resource constraints and use case, we find a closer alignment with grassroots community data initiatives, where grassroots communities self organize to collect and make use of their data for social or political causes, often using mainstream platforms and technologies. For example, in Quotidian Report, citizens in Mexico report crime and local incidents on Facebook groups to generate aggregated data on public safety [1]. Similarly, the 996.ICU initiative[3] involves Chinese IT workers sharing their work schedules on a GitHub repository to protest against long working hours. Both initiatives successfully mobilized and sustained community participation, leveraging data contributed by community members to address issues that were otherwise overlooked or suppressed. Taking a similar approach, the StammerTalk community not only produced a sizable, representative, and versatile speech dataset to address their unmet technological needs, but also enhanced capacities and connections within their community through data collaboration. Nevertheless, questions remain regarding the legal framework and maintenance mechanism for the cocreated dataset, and we hope our work serves as a placeholder for future investigation into community-led data models for grassroots and underserved communities.

Overall, our work contributes to the ongoing efforts in building fair and inclusive speech AI for stuttered speech. Our contribution lies not only in introducing the first dataset of stuttered speech in Mandarin Chinese but, more importantly, in envisioning a new, sustainable partnership between the AI community and the disability community in data collaborations that address fairness challenges faced by people with disabilities.

## 3 BACKGROUND

Here, we provide an overview of the StammerTalk community and its members as background information for their data collection initiative. We also describe the procedure and steps of data collection and the activities and roles taken by community members involved in this process. The information presented was sourced from public channels, such as StammerTalk's public account on WeChat and podcasts, as well as our conversations with community members. Finally, we disclose our relationship and the mode of interactions with StammerTalk community beyond this research in our positionality statements, discussing potential influence and power dynamics from our own identities and backgrounds. Note that the data collection process is the subject of our study, but not the study itself. This paper performed meta analysis of the data collection process, and we will describe our methods in Section 4.

### 3.1 StammerTalk Community

StammerTalk (口吃说) is an online community for Chinese-speaking people who stutter. Started in early 2020 as a podcast featuring interviews with and by people who stutter, it now runs a variety of advocacy, education, and community programs, including: 1) a WeChat public account sharing personal stories and research findings on stuttering; 2) a WeChat group for Chinese speaking individuals who stutter; 3) bi-weekly virtual self-help groups; 4) large community events, such as an annual virtual conference on International Stuttering Awareness Day. Through these programs,

[2]https://driversseat.co/
[3]https://github.com/996icu/996.ICU

Table 1. **Background Information of StammerTalk Core Team Members**

| Name | Gender | Age | Country | Occupation | Community Role | Data Collector |
|---|---|---|---|---|---|---|
| Rong | M | 25-35 | Austria | Research scientist in a large technology company | StammerTalk co-founder | Yes |
| Lezhi | F | 25-35 | U.S. | Data scientist in a large retailer company | StammerTalk early member | Yes |
| Jia | F | 25-35 | U.S. | Ph.D. student in Communicative Sciences and Disorders | StammerTalk co-founder | No |

the community has grown to include hundreds of members in its WeChat group and following its public account, with an average of around a hundred participants attending to its virtual conferences. To the best of our knowledge, StammerTalk is one of the largest communities for Chinese-speaking people who stutter.

Despite its size and success, StammerTalk operates entirely as a grassroots community in virtual spaces. Its membership is informal, fluid, and geographically distributed, with no formal process, fees, or mandatory participation in its events and activities. As a result, StammerTalk does not have a formal budget, full-time staff, or legal status in any country, but relying on the dedication of its volunteers. A team of ten community members volunteered to mainly daily tasks like hosting self-help groups, content production, and event management. Collaboration among volunteers is loosely-structured and flexible, with a "core team" of three members provide leadership and direction. Please refer to Table 1 for more information on their backgrounds. Operational tasks are allocated based on individuals interests, skills, and availability. The volunteers coordinate through only online channels, such as WeChat groups and video calls.

In summary, StammerTalk is a grassroots community led by and for Chinese-speaking individuals who stutter. With its members predominantly reside in China, a region where stuttering stigma is more profound and professional support is much more limited comparing to western societies [20]. It provides a unique space for Chinese-speaking people who stutter to find community and learn more about stuttering, despite having limited resources.

### 3.2 Stuttered Speech Collection Process

StammerTalk's efforts to create the stuttered speech dataset spanned over one year period, taking several important steps from project conception, preparation, participant recruitment, speech recording, and speech annotation. We detail these steps below.

*3.2.1* **Conception.** The idea of creating a Chinese language stuttered speech dataset emerged in a WeChat conversation between the StammerTalk core team and one author of this paper in December 2022. Recognizing the lack of a representative stuttered speech dataset in Chinese language, StammerTalk's core team saw the opportunity to co-create such dataset as a valuable resources to improve ASR services for Chinese stuttering community. One of the core team members, Rong[4], volunteered to lead this initiative.

*3.2.2* **Preparation.** Before kicking off the data collection, StammerTalk core team carefully planned the process and located resources and partnerships they needed. They pitched the project to a wide range of individuals and organizations and established partnership with prominent fluency researchers, AI researchers, US-based nonprofit organizations, and

[4]Throughout this paper, we use the real names of StammerTalk community members whenever possible with their explicit permission.

a AI data service company in China. These partnerships enabled StammerTalk to develop comprehensive technical specifications for their data collection, build rigorous and AI-friendly annotation guidelines for Chinese stuttered speech, access legal services, and receive free annotation services with the collected speech data. In particular, significant amount of time and efforts were spent with Chinese, EU, and US technology law specialists to draft participant agreements that maximally satisfied the data regulations and compliance in different regions.

*3.2.3* **Participant Recruitment.** Participants of the data collection were recruited on WeChat through StammerTalk's public account. The first recruitment message was posted in January 2023. The message emphasized the objective of the data collection to improve speech AI for stuttered speech, and introduced the basic process and compensation (¥100 RMB ($14 USD) cash via WeChat pay and a swag from the speech annotation partner) for participation. The recruitment was deliberately made open to anyone self identified as a person who stutters, without restrictions on age, gender, or stutter severity. Interested participants were directed to Rong to schedule.

The first recruitment successfully attracted over 40 interested participants within a few days. After completing the data collection with participants from the first recruitment, a second recruitment with the same message was run in July 2023, leading to another 30 participants.

*3.2.4* **Speech Recording.** Upon signing up for a data collection session, interested participants would receive a participant agreement form for them to review. This form detailed the purpose of the data collection, potential applications of the collected data, privacy protection measures, and opportunities for participants to be involved in data management. Once the form was signed, interested participants were scheduled for a 60-minute data collection session with the interviewer (i.e., one of the two StammerTalk core team members, Rong or Lezhi) who also stutters via Zoom or Tencent Meet, structured as follows:

(1) **Introducation (5 mins)**: The session started with an self introduction by the interviewer. The interviewer then briefed the participant on the recording tasks and activities. Additionally, the interviewer checked the technical and environmental setup of the interviewee to ensure audio quality.
(2) **Unscripted Spontaneous Conversation (30 mins)**: The interviewer led a casual conversation with the participant, with topics around the participant's personal background and lived experiences with stuttering.
(3) **Voice Command Recitation (30 mins)**: Participants were provided a set of common voice commands to read aloud. These commands were designed and curated by the group of partner researchers mentioned in Section 3.2.2, and covered a wide range of topics, such as control commands for smart home devices, names of music tracks, films, news headlines, and locations. The commands were typically short, ranging from 1 to 64 (median of 8) characters. This selection was made to ensure a comprehensive representation of everyday voice commands. An example of these commands is "你好，米雅，这首歌循环六遍" (Translation: "Hello, Miya, repeat this song six times").

The latter two components of the session were audio recorded locally in the interviewer's computer. Subsequently, these recordings were uploaded to a shared Google Drive folder, accessible only to the StammerTalk core team and selected partners for further processing. Approximately an hour of speech data was collected from each session.

*3.2.5* **Speech Annotation.** Given the absence of guidelines for annotating stuttered speech in Chinese, Rong extended existing annotation guidelines for fluent speech with stutter-specific instructions adopted from similar work in English [25]. He also sought inputs from SLP researchers and other PWS. The guidelines were refined through three iterations, each with a trial run with professional speech annotators who do not stutter. Rong also provided necessary

feedback and training for the annotators after each trials to help them better identity, annotate, and transcribe stuttering events. The trained annotators performed the the speech-to-text transcription and stuttering event annotation for all the speech recorded in the data collection sessions.

By December 2023, a total of 70 people who stutter (not including the interviewers) had participated in the data collection process. After consulting with their technical partners, the StammerTalk core team decided to publish the 70-hour dataset first for technical explorations, before collecting more data.

### 3.3 Positionality Statement

Recognizing that as researchers, our personal backgrounds and identities shape how we engage with communitiesx and interpret our findings, we outline our backgrounds and perspectives below.

Both of us are Mandarin-speaking, Asian/Asian American women residing in North America. Together, we bring 22 years of experience working in academia and the corporate, with expertise in data science, HCI, accessibility, and AI. While affiliated with technology companies and/or university research institutes, we both had experience gathering data from individuals with disabilities, either directly through company's or institution's platforms, or indirectly via data vendors. One of us identifies as a person who stutters. This author has engaged with StammerTalk, attending a self-help session and being interviewed for their podcast. Additionally, she has personal and professional ties with the StammerTalk moderators through other stuttering-related advocacy and technical projects.

Though our close relationship with StammerTalk and shared experiences as stutterers brought trust and community access, it didn't entirely negate the power dynamics between researchers and subjects. Our socioeconomic and educational backgrounds also granted us certain privileges relative to many community members we engaged with.

## 4 METHODS

To understand the process, benefits, and challenges of this community-driven stuttered speech data collection led by StammerTalk, we conducted **semi-structured interviews** with the primary data collectors to explore their motivations, experiences, and challenges. We also developed and administered a **survey** to the data contributors, further gaining insights into their perspectives. These methods, detailed below, were designed to capture a holistic view of the initiative, exploring both the experiences of those leading the data collection and the perspectives of those contributing data. This comprehensive approach allowed us to gain in-depth insights into the entire data collection process and its broader implications.

To distinguish participating community members with different roles in the initiative, for the rest of this paper, we will refer to the StammerTalk core team members who collected and processed the data as **data collectors**, and the community members who signed up to participate in the recording sessions as **data contributors**.

### 4.1 Semi-structured Interview with Data Collectors

We conducted semi-structured interviews with the two primary data collectors of this initiative. Our goal was to delve deeper into their motivations, capture their experiences, and understand the challenges and insights they garnered as leaders throughout the data collection journey. As detailed in the background, StammerTalk operates as a grassroots organization heavily reliant on volunteer efforts, resulting in limited resources. Consequently, all recording sessions were conducted by these two moderators. Each moderator had conducted interviews with approximately 30 data contributors at the time of this study, providing them with a wealth of experience. This extensive involvement ensures that they

could offer comprehensive and in-depth insights, making their contributions particularly valuable and representative for our research objectives.

**Interview Procedure.** One of the authors conducted the remote, semi-structured interviews via Zoom. With the consent of the two data collectors, each session was audio-recorded and later transcribed verbatim. The duration of both interviews are 80 and 90 minutes, respectively. Both data collectors volunteered for the interview without receiving any monetary compensation. The names and background information of the two data collectors can be found in Table 1. Per the preference of the data collectors, we will refer them with their real names.

**Interview Protocol.** The interview process was meticulously structured to cover various aspects of the data collectors' experiences. It comprised several key segments, each focusing on different elements of their involvement and reflections:

- **Warm-up Session:** Data collectors share about their professional roles and describe personal experiences and challenges related to stuttering.
- **Motivation and Incentives:** Asking data collectors about their inspiration or driving force behind participating in the initiative.
- **Processes and Experiences:** Detailed exploration of preparation, planning stages, execution of tasks, and handling deviations and unforeseen circumstances. Discussion includes distribution of responsibilities, technical setup, participant recruitment strategies, anticipated workloads, and timelines, as well as any deviations from the initial plan and lessons gleaned from the overall process.
- **Challenges and Strategies:** Data collectors reflect on anticipated and unexpected hurdles and strategies employed to overcome them.
- **Introspection:** Prompting data collectors to introspect on their journey, emphasizing lessons learned, personal growth, and future plans. Offering an open platform for sharing additional insights or anecdotes.

**Interview Analysis.** We used an inductive thematic analysis process to analyze the interviews. First, Two authors independently reviewed the interview transcripts to identify salient ideas and patterns. Utilizing these insights, they developed an initial codebook that encapsulated primary and secondary themes emergent from the data. Both authors then engaged in a thorough discussion, comparing and contrasting the themes they had individually identified in collaborative sessions. Through a process of deliberation and synthesis, overlapping or closely related themes were merged to ensure clarity and coherence. We present our themes and results in the following section. Both interviews were conducted in Mandarin, participant quotes are translated to English by the authors and reviewed by Rong and Lezhi.

### 4.2 Survey with Data Contributors

Our initial interviews with the data collectors yielded valuable insights into the data collection processes, and the unique challenges and dynamics encountered in moderating interviews with people who stutter. These narratives significantly informed our preliminary research questions. Additionally, StammerTalk had implemented a brief exit survey, including a 5-point rating scale for assessing data contributors' experiences and an option for additional comments. Conducted at the end of the recording sessions, this exit survey captured the immediate reflections and experiences of the data contributors.

In pursuit of a more comprehensive perspectives from the data contributors, we expanded our methodology to incorporate an extensive survey targeting the data contributors. This expansion, aimed at enriching the themes identified

in the moderator interviews, was informed by both the initial interviews and the exit survey responses. While also serving to validate these themes, our primary focus was on broadening and deepening our insights. The survey questions, predominantly of a 'select-all-that-apply' nature, were designed to capture a diverse range of experiences and perspectives from both the data collectors and data contributors. This methodological expansion was integral in capturing a holistic view of the data collection process and its nuances. The survey was conducted in Mandarin, and the results are presented in subsequent sections in English translation by the authors.

**Survey Questions.** The survey comprised 14 distinct questions, both open- and closed-ended, categorized into the following segments:

- **Demographics**: This section gathered data on respondents' age, gender, occupation, and previous stutter-related support or interventions they might have received.
- **Reasons for Data Contribution**: This section sought to understand participants' motivations for joining the data collection initiative. It employed the maximum difference scaling method to discern the intensity and preference of their motivations.
- **Overall Experience**: Here, participants rated their overall experience through a Likert scale. Follow-up questions then delved into specific factors that either enhanced or detracted from their experience.
- **Evaluation of the Interviewer**: Participants were prompted to assess the interviewer using a Likert scale. Subsequent questions sought feedback on the interviewer's strengths and areas of improvement.
- **Challenges**: This section was dedicated to understanding any obstacles or challenges participants faced during their data collection interview.
- **Engagement with StammerTalk**: Participants were queried about their past engagements with StammerTalk activities and whether they'd be inclined to participate in future initiatives hosted by the organization.
- **Personal Takeaways**: An open-ended section, this allowed participants to articulate what they perceived as their most significant gain from the entire process.

Through this structured approach, the survey was designed to comprehensively capture data contributors' experiences, challenges, and insights. For a comprehensive view of the entire survey, please refer to the Supplementary Material.

**Recruitment.** Data contributors were individually invited by Rong, one of the data collectors, to complete the survey. They were informed that the survey was administrated by [Organization Name], designed to better understand and improve the data collection process, and they were be compensated with ¥30 RMB (approximately $5 USD) upon completion of the survey. The survey was hosted through Tencent Survey platform. The survey took about 5 minutes per respondent, and compensation were distributed by Rong on behalf of [Organization Name] to the respondents through WeChat Pay.

**Analysis.** Among all the 58 data contributors who completed the data collection sessions by the time we administrated the survey, 55 people (95%) submitted their responses to the survey. The mean survey completion time was 5 minutes.

For open-form questions, we utilized an iterative coding methodology [18]. For each question, one author developed an initial codebook. Two authors then collaboratively discussed and refined the codebook, applying it iteratively to all responses. To analyze the quantitative data, we focused on descriptive statistics, primarily frequencies. Given the nature of our survey, which aimed to understand holistic experiences rather than identifying correlations between variables, most questions were of the "apply-all-that-apply" type. Thus, complex statistical analyses were not deemed appropriate or necessary for our research objectives.

**Participants.** Of all 55 respondents, 17 individuals (30.9%) are 18-24 years old, 31 (56.4%) are 25-34 years old, 6 (10.9%) are 35-44 years old, and 1 (1.8%) is 45-54 years old. The majority (63.6%) of the survey participants identified as male, while the other 20 people (36.4%) identified as female. Our data contributors have a wide range of occupations: a significant number of participants (23.6%) identified as students; other notable occupations include IT-related roles (11%), medical professionals (7%), public service roles (e.g., civil servants, teachers), and roles in various specialized fields ranging from energy sectors to biotechnology.

The majority of our participants (83.6%) also indicated that they have received some form of stutter-related support in the past, with the types of support not being mutually exclusive. Specifically, 25 participants had undergone stuttering therapy or training, 27 had attended online or offline stuttering self-help groups, another 27 identified as members of online or offline communities for people who stutter, such as the StammerTalk WeChat group or National Stuttering Association (NSA) in the U.S., and 17 had participated in stuttering-related community events like lectures or public activities. Conversely, 9 individuals (16.4%) reported not having engaged in any of the aforementioned forms of support.

## 5 FINDINGS

Here we describe the major findings from our work, centering around the incentives, experiences, gains, and challenges for community members to lead and participate in the data collection process. Our findings highlight that, contrary to what is reported in previous research [29], StammerTalk members who participated in the community-led data collection were driven by intrinsic incentives - such as the making meaningful contribution to the community and connecting with other community members, rather than monetary compensation. Community members also gained empathy, understanding, knowledge, and personal connection with each other during the data collection, resulting in overwhelmingly positive experiences and a sense of self and community empowerment.

Our data also uncover the challenges for community-led data collection, namely, the significant time commitments, the resources required to annotate the recorded speech data, and the uncertainties with legal and privacy implications. While the StammerTalk community was pragmatic and resourceful to navigate these challenges, our study calls for the development of adequate socio-technical infrastructure for a broader and easier adoption of community data stewardship model from other marginalized communities.

### 5.1 Incentives

The StammerTalk community's primary drive for the stuttered speech collection project stemmed from **intrinsic motivations** such as community empowerment and forging interpersonal connections, overshadowing external incentives like monetary rewards.

Both data collectors, Rong and Lezhi, have backgrounds in technology and felt compelled to contribute their skills to address the community's technological challenges. Rong, who works at a speech technology company, shared that, "*I'm professionally involved in this space, understanding the entire process well. (...) Therefore, undertaking this project end-to-end would be very meaning for me.*"

Their stuttering and technical background also enabled Rong and Lezhi to quickly recognize the dataset's potential impact on stuttering specific research, education, and technologies, especially in the Chinese language context. For example, Rong expected that " *such stuttered speech dataset would not only benefit the research and development of (speech AI) technology, but also, for the training of Speech and language pathologists (SLPs) (...) it could be very helpful.* "

Additionally, Rong and Lezhi also saw this project as a potential asset for their careers. Rong, already working speech technology R&D, considered leading the project end-to-end, starting from data collection, as a valuable professional

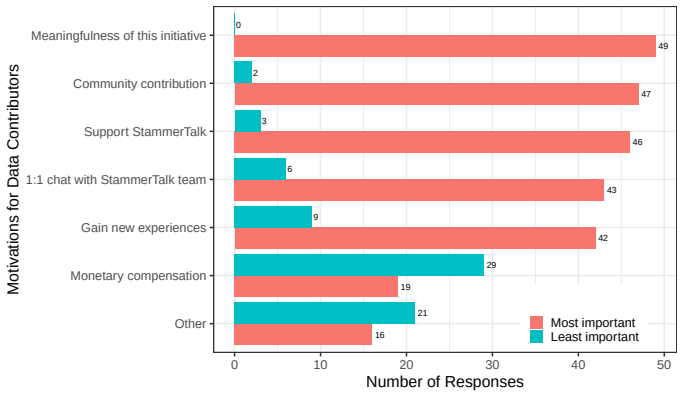


Fig. 1. The most and least important reasons for data contributors to participate in data collection project.

experience. Meanwhile, Lezhi believed that listing a project like this on her resume would empower her to more easily disclose her stuttering and distinguish herself with co-workers, managers, and potential employers. Both Rong and Lezhi viewed the data collection project as an act of self- and community advocacy. As Lezhi elaborated,

> *I want to publicize my stutter... I want to empower myself through stuttering. (...) I want to differentiate myself from others, from people who do not stutter. What's my advantage? My longstanding involvement with the stuttering community gives me insights into the unique challenges faced by stutterers. (...) This equips me well with ideas on leveraging technology to improve experiences of people who stutter, especially since current technologies often overlook their needs.* (Lezhi)

Echoing the sentiments of the data collectors, most data contributors' participation in the data collection were not driven by material gains, but their recognition of the value of this project to the stuttering community and their desire to contribute to and engage with the community. As shown in Fig. 1, when asked to pick the most and the least important reasons for them to participate in data collection, more than 80% of the 55 survey respondents found their top motivators to be: the innate value of this project ("*meaningfulness of this initiative*", N=49), contributions to the stuttering community ("*community contribution*", N=47), support for StammerTalk ("*support StammerTalk*", N=46), opportunity to talk to other PWS ("*1:1 with StammerTalk team*", N=43), and opportunity to gain new and interesting experiences ("*Gain new experiences*", N=42). While the motivations like the impact of data and the willingness to contribute to the community were also reported in previous research [29], the desire to support the data collection organization (StammerTalk) and to interact with the data collectors (StammerTalk team) are novel and interesting, highlighting the value of the existing reputation of StammerTalk team and the personal connections within community members.

On the other hand, a relatively small number (N=19/55) of the survey respondents rated "*Monetary compensation*" as the most important reasons to participate. In fact, consistent with previous results [29], "*Monetary compensation*" was the frequently picked (N=29/55) as the least important reason(s) to participate in the data collection. Last but not least, eight people out of 16 who selected "Other" and provided the description, were mostly elaborating on reasons of "Community contribution" (e.g. "support all activities related to stuttering") and "1:1 chat with StammerTalk team" (e.g. "develop the courage to communicate with strangers.").

To sum, the StammerTalk community were intrinsically motivated to conduct and participate in the data initiative. Leveraging existing technical talents within the community, community members contributed their speech data to

make a meaningful contribution to the community, address their needs and rights, build deeper connections with each other, and embrace their - often marginalized - identity as people who stutter.

### 5.2 Experiences

While previous work highlighted the heightened stress and "performance anxiety" for people with disabilities during data collection tasks [29, 43], participants of the StammerTalk community data collection found their experience during the data collection highly satisfying and enjoyable.

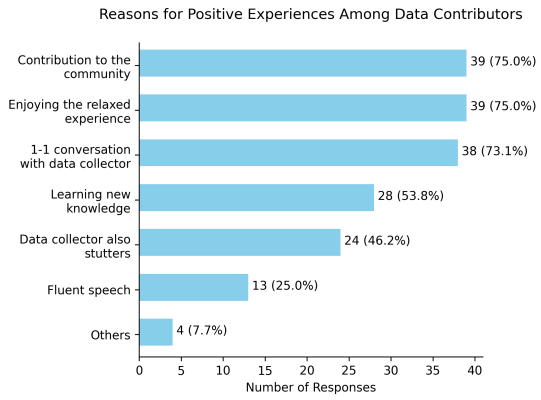


Fig. 2. The primary reasons that led to the positive experiences among the data contributors in the data collection project.

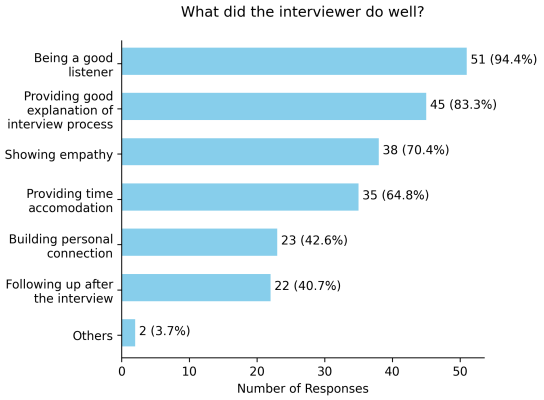


Fig. 3. Data contributors' feedback on data collectors' competencies during the data collection project.

The vast majority (95%, N=52/55) of the respondents described their experience with the StammerTalk team's recording session as either "*Very satisfying*" or "*Satisfying*". Those who reported a positive experience were prompted to pick the primary factors contributing to their feelings, and the data is summarized in Fig. 2. The three leading reasons contributing to the positive experiences of data contributors were: a sense of making a meaningful contribution to the community (75%, N=39/52), a relaxed and comfortable atmosphere during the interview (75%, N=39/52), and the unique experience of having a one-on-one conversation with another person who also stutters (73%, N=38/52). These results resonate with our earlier findings regarding the primary motivations for participation, confirming the value of stuttering community and the connections with other PWS for the data contributors.

While previous research reported that the inaccessibility of the data collection environment and process could create significant physical and psychological stress for participants with disabilities [29], our results highlight the stark difference in community-led data collection: StammerTalk's data collection sessions were a source of pleasure and enjoyment, rather than exhaustion or stress.

The data collectors played an important role in making the data collection session pleasant for the data contributors: majority of data contributors found their interaction with data collectors during the data collection process uniquely positive, greatly contrasting with their typical speaking experiences. Of 55 respondents, 54 rated their interaction with the data collectors as either *Good*" or *Very Good*." As shown in Fig.3, respondents particularly valued the data collectors' attentive listening (94%, N=51/54), clear communication about the data collection process (83%, N=45/54), and the substantial empathy shown by the interviewers (70%, N=38/54).

A significant number of data contributors (43%, N=23/54) particularly enjoyed being interviewed by someone who also stutters. As Rong observed, the mutual experience of stuttering established an immediate sense of trust. He recalled the participants often remarked, " *oh, you also stutter!', followed by, 'now I can relax.'* " Lezhi's observations resonated, " *People who stutter usually engage in a psychological defense when it comes to speaking,(...) Since my stuttering is relatively severe, the participants might feel there is nothing they need to hide when speaking with me.* "

To achieve a mutually positive experience, the data collectors also adopted thoughtful and respectful communication behaviors. They shared personal experiences with stuttering, adapted conversation topics to accommodate the participant's speech and emotional state, and showed genuine interests and patience towards what the participant had to say. For example, Lezhi noted her ability to anticipate and sense the discomfort and accommodate accordingly:

> *When someone was nervous, I would chose to ask them some easy topics to help them relax. (...) As a person who stutters, I know what types of topics will make them more nervous, I could also quickly identify the characteristics of their stutter and which words might be difficult for them to say.* (Lezhi)

Consequently, the supportive communication environment enabled some data contributors to speak more fluently than usual during data collection, showing less stutters in their speech. As it is not uncommon for PWS to find stuttering uncomfortable and prefer speech fluency [11], 13 out of 54 survey respondents did cite the increased fluency as a factor in their positive experience. However, the boosted fluency could result in the divergence of the recorded speech from people's typical stuttering patterns, creating a potential challenge to the representativeness of the dataset.

### 5.3 Gains

Beyond the direct, tangible benefit of creating a data asset for the community, data controllers and data contributors also gained valuable skills, experiences, knowledge, and connections that could lead to long-term efficacy of the community.

*5.3.1* **Data Collectors: Personal Growth, Broadened Perspectives, Relationships.** While neither Rong nor Lezhi received any monetary rewards from working on this project (Rong even spent personal funds to compensate particpants), they identified personal growth in several areas, including 1) enhanced interpersonal communication skills, 2) strengthened bonds within the stuttering community, and 3) a more comprehensive understanding of the diverse personal and social contexts surrounding stuttering.

Both Rong and Lezhi had evolved as listeners and conversationalists over the course of the data collection process. Reflecting on his journey, Rong remarked:

> *I learned a lot (from conducting the interviews). I learned how to listen, especially to someone who stutters, (...), and to keep the conversation fluid. (...) They (people who stutter) wanted to have a real conversation with you. Initially, I was a bit rigid. But after receiving feedback, I improved the way I posed questions and showed genuine interest in their life stories. This way, the interview experience became much better.* (Rong)

Rong and Lezhi also appreciated the opportunity to interact with PWS from diverse backgrounds and gain broader perspectives on stuttering. Lezhi reflected, " *Beyond the project's tangible outcome, the true reward was engaging in discussions with numerous people who stutter and absorbing their varied viewpoints.* "

The relationships cultivated between the data collectors and contributors were not transient but of lasting values. Both Rong and Lezhi maintained personal connections with many data contributors post-data collection through social platforms like WeChat.

*5.3.2* **Data Contributors: Unity, Acceptance, Knowledge.** Our analysis revealed that for the data contributors, the primary gain from participating in the data collection was not the monetary compensation they received ($14 USD), but rather the sense of unity, self-acceptance, and a deeper understanding about stuttering. These benefits align with, and even surpass, their initial motivations for participation.

Many data contributors (N=21) shared that participation in the data collection project strengthened their feelings of unity, recognition, and empowerment within the stuttering community, fostering a deeper sense of belonging and collective power. One data contributor expressed (P19), " *[I love] meeting more friends and teachers. It made me realize that there are many people in the world just like me. We all strive to live well, working hard to overcome the impact of stuttering on ourselves.* " Others (e.g. P20), acknowledged the broader awareness and understanding brought about by the project to the general public: " *I realized that there are so many people continuously paying attention to the stuttering community... leading more people who stutter to focus on themselves.* " This growing unity and recognition, as summarized by another participant P1, has led to a feeling that " *our community has united and received more attention, advancing the progress of stuttering treatment in China.* "

Data contributors (N=14) also highlighted the immense personal growth, realization of their inherent potential, and emotional relief gained from the genuine, one-on-one conversations with other people who stutter. Free from judgment and without the burden of hiding their stutter, they felt a profound sense of liberation and empowerment. Engaging with someone from " *a similar group* " deepened this transformative experience, accentuating the power of shared experiences and the realization of one's true potential. As P33 expressed, being able to " *freely express without consciously hiding my stutter* " not only served as a medium of self-expression but also as an affirmation of self-acceptance and self-worth. The understanding and respect they gained, especially from an interviewer who also stutters, instilled a sense of hope and a more positive attitude in life.

Other data contributors (N=10) say that the biggest gain from participating the data collection project is having learned new knowledge about stutter. For instance, P9 mentioned " *I learned that one can approach stuttering from a scientific perspective.* ". Others emphasized the learning gained uniquely from talking to people who also stutter. As P45 put it: " *The interviewer's pronunciation and manner of speaking in a very slow and gentle voice slightly improved their speech fluency [...] This deeply resonated with me, and I am currently learning this way of speaking.* "

In summary, the data contributors greatly valued their participation in the data collection project as it left them with a stronger sense of community, self empowerment, and new knowledge on stuttering. Similarly, data collectors experienced personal growth and formed lasting connections. While previous research rarely studied the perspectives of data contributors post data collection, our findings showed the community-led data collection's profound positive impact beyond its primary objective, highlighting its promise as a healthy and beneficial model for collecting AI data from the disability community.

## 5.4 Challenges

Despite the community members' strong motivation and positive experiences, some substantial challenges are unavoidable during the process. While the StammerTalk community had managed to come up with creative strategies to navigate these challenges, some questions remained open as the project moves forward.

*5.4.1* ***Challenges for Data Collectors.*** Data collectors faced four major challenges as summarized as follows:

***(1) Time Commitment:*** Rong and Lezhi, both full time IT professionals in Austria and the United States, dedicated their evenings and weekends to the project. The time zone differences between data collectors and many of the

participants in China led to scheduling challenges. This limited time and schedule availability, coupled with unforeseen last-minute reschedule requests from participants, resulted in a maximum of one or two recording sessions per week. The data collectors were mentally prepared for such operational overhead, Rong anticipated that the extensive time required for recruitment and scheduling could extend the project's timeline significantly, possibly over a year, to achieve the target of 100 hours from 100 individuals.

The data collection sessions were time consuming, too. As Lezhi recalled, many participants enjoyed the conversations so much that their sessions went significantly over time. In those situations, she would guide the participant to finish the planned speech tasks first and continued the conversation after completing the recording. The time intensity of the data collection process has been a major challenge for the StammerTalk team, especially, when the workload was split by only two volunteers - Rong and Lezhi - using their spare time outside demanding IT jobs. Rong had called for other volunteers with the StammerTalk community as data collectors, but did not receive any responses. The demanding time commitment also contributed to an early stopping of the data collection, after having only 70 participants rather than the planned one hundred. Better tools and recognition for data collectors could help alleviate the time intensity and reduce individual data collectors' workload.

***(2) Data Annotation.*** As briefly introduced in the *Background* section, finding annotation services to accurately annotate the collected Chinese stuttered speech sample was also challenging, as it had never been done before at this scale. As a result, Rong had to spent substantial amount of time and energy to create detailed annotation guidelines and to train the annotators, who were non-stuttering and had no prior experience of annotating stuttered speech. While some existing stuttered speech datasets skip transcribing stuttered utterances (e.g. [23]), Rong made the deliberate decision to transcribe stutter verbatim, so that stutters are authentically represented rather than erased. However, this decision did increase the difficulty and workload for the annotators. For example, the annotators had a hard time detecting all stuttering events or differentiating natural disfluency vs. stuttering disfluency. It took three iterations for the annotators to be able to identify and label the stuttering events correctly. During each iteration, Rong would carefully review the annotations produced by the annotators, and returned with corrections with detailed explanations. At the end, he also carefully reviewed and verified all annotations and transcriptions to ensure the accuracy and completeness of the dataset. Although the entire process was tedious and time consuming, Rong recognized the dedication of the annotators and their adaptability, but also realized that, due to the pro bono nature of the service, achieving the ideal annotations consistent with stuttering professionals was ambitious:

> *It took the annotators quite a lot of efforts during our training. Since none of them stutters, nor did they work with PWS professionally, it is very difficult for them to produce the consistent annotations as stuttering professionals do. After three iterations, although there were still some places that were unsatisfactory to me, I thought it was already very good for non-stuttering annotators to have this level of quality in their annotations.* (Rong)

***(3) Data Quality and Representativeness.*** Another key challenge faced by the data collectors was ensuring both the quality and representativeness of the recorded speech. They aimed to balance between capturing clear sound, diverse speech types, and varying stuttering patterns, sometimes at the cost of the positive experience of the data contributors.

Concerning **sound quality**, although data contributors received guidelines on environmental and technical settings, not all complied. For instance, Lezhi encountered situations where contributors were in noisy surroundings or interrupted by phone calls, necessitating either waits or rescheduling to achieve optimal sound conditions.

The data collectors also strove to have the data sufficiently cover **the variety in stuttering patterns and severity levels**. Stuttering, similar to many other neuro-developmental conditions, varies in frequency, severity, and manifestation across individuals and contexts [24, 43]. The recording sessions –combining unscripted conversations with recitation of common voice commands – aimed to capture different speaking contexts. However, the comfort ambiance often led to participants stuttering less than usual, particularly during voice command recitation, which could limit the data's real-world representativeness.

To address this issue, the data collectors employed strategies, such as 1) encouraging voluntary stuttering – imitating stuttering on words they typically would not stutter on, and 2) posing challenging questions to induce tension.

While these strategies help increase the frequency of stuttering, there are trade-offs, such as the tradeo-off of tension and openness during the unscript conversations. As Lezhi explained,

> *There needs to be a balance. When someone was nervous, they could choose to speak less; when someone was relaxed, they would not stutter. When someone was nervous, I would chose to ask them some easy topics to help them relax; when someone was very relaxed, I would ask a less comfortable question. As a person who stutters, I know what types of topics will make them more nervous. (...) Based on what he (the data contributor) shared about his background, I would intentionally follow up with some additional questions make him feel like at a job interview, to create a bit more tension.* (Lezhi)

Despite the lower-than-expected stuttering frequency, the data collectors believed their method best represented and empowered the stuttering community. Data contributors were not pre-screened to participate. While they did complete the Overall Assessment of the Speaker's Experience of Stuttering (OASES) [45], it was not used as a selection criterion but rather as metadata. Rong reflected upon the recruitment process, and emphasized that a person's self-identification as someone who stutters should be the sole requirement for participation to avoid external biases. This approach accentuates the difference between community-led and expert-led data collection. *Unlike commercial entities that might exclude someone for not being "disabled enough", community-led efforts, like this one, prioritize self-identity and inclusion.*

***(4) Data Protection and Governance.*** Ensuring data protection and governance posed a another notable challenge. Given that interviews delved deep into contributors' stuttering experiences, many participants shared intimate details about their lives. Such openness enhanced the data's authenticity but also raised privacy concerns. Though data collectors attempted to safeguard privacy by editing out personal details and encouraging pseudonyms, the effectiveness of these measures in fully anonymizing the dataset remains uncertain. While contributors consented to sharing their data for non-commercial uses, uncertainty around whether and how to share this dataset remains even among the community itself. Rong supports releasing the dataset under a non-commercial license, while Lezhi, citing legal and privacy concerns, believes only analyses and models derived from the data should be open-sourced. The complexities of global regulations surrounding biometric data, which includes speech, coupled with limited resources and expertise of StammerTalk being a grassroots online community, introduce potential legal liabilities for data collectors.

***Challenges for Data Contributors.*** As shown in Figure 4, of the 49 data contributors who responded to the survey question, 18 (36.7%) identified their stuttering during the interview as a challenge they faced. Another 13 (26.5%) participants felt nervous, an emotion that aligns with findings from the previous "Experiences" section where many contributors revealed they were still self-conscious about their stutter, and many desired fluency. This sentiment mirrors the often-felt physical tension and discomfort that many who stutter experience during speech. In a contrasting vein, 17 (34.7%) participants found it challenging to deliberately stutter during the interview, a finding that intriguingly mirrors the challenges faced by data collectors who sought to elicit a broader range of stuttering for data diversity. 16

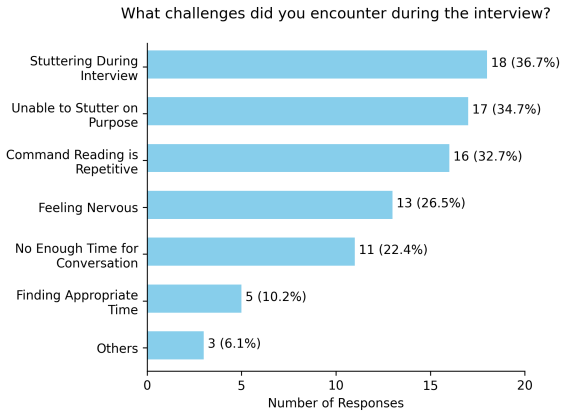


Fig. 4. Primary challenges faced by data contributors during the data collection process.

(32.7%) participants found the voice command reading tasks monotonous, especially since they had to repeat several commands multiple times. Additionally, 11 contributors expressed a desire for more 1-on-1 interaction time with data collectors, underlining their interest in engaging and learning from community leaders and members.

To sum, the key obstacles we identified in community-led AI data collection for PWD include: time, labor, and legal resources, legal and privacy uncertainties, and finding the right balance between accurately capturing the characteristics of disability and the discomfort experienced by the data contributor.

## 6 DISCUSSION

### 6.1 Comparison Between Community-led and Expert-led AI Data Collections

Our findings highlight several differences between grassroots community-led AI data collection and expert-led efforts.

*6.1.1* **Agency.** The StammerTalk community conceptualized, planned, and executed the data collection process with full agency and autonomy. Distinct from expert-led, commercial data collection (often by technology companies or research institutions) in which the participation was often driven by monetary compensation [29], StammerTalk's data collection, originated from the community's own needs and goals, drew substantial interests and participation with only a modest compensation promised ($14 USD per hour). The community data collectors also had the full autonomy to design the data collection procedure and objectives, maximizing community values such as inclusion and acceptance. For example, actively rejecting the "medical model of disability" that disabilities were defined by medical experts and authorities [11], the community chose to include anyone *self-identified* as a PWS in their dataset, without a screening or qualification process that is commonly implemented in expert-led data collections [28]. Similarly, to normalize stuttering and push back on AI's embedded expectation on fluency today [24], the community made the call to transcribe stuttered utterances verbatim, despite its additional annotation costs.

*6.1.2* **Authenticity.** The StammerTalk community was able to represent themselves authentically in their data. Stuttering is known to be highly variable: the severity of stuttering can vary significantly across individuals, environments, and conversation partners [39]. As a communication disorder, stuttering is inherently social: most PWS do not stutter when they are alone [11]. Given the nature of stuttering, conventional speech data collection method, in which the

speakers record monologues with given prompts [28, 29], works poorly in capture authentic, real-world stuttering behaviors. StammerTalk data collection included spontaneous, unscripted conversations between two people who stutter in a natural setting, a type of speech samples that are highly valuable but not yet available to AI models. The recorded conversations were also designed to cover topics and personal stories related to stuttering, encouraging authentic and open expression of the data contributors of their otherwise stigmized identity as a PWS and fostering general awareness and empathy for stuttering in the AI research community.

In [a recent paper] (citation redacted for anonymity), we confirm that the StammerTalk dataset captures the variability and heterogeneity of stuttered speech through descriptive analysis. Specifically, the dataset illustrates the variation in disfluency rates across different tasks and speakers, reflecting the dynamic and situated nature of stuttering. For instance, while participants stutter more in Conversations (mean=9%) than in Voice Command Dictation (mean=7.1%), the frequency of stuttering varies more in Command Dictation (std=0.15) than in Conversation (std=0.08). Furthermore, the dataset highlights the diversity in stuttering patterns among the 70 participants: some might speak with more word repetition, while others experience more blocks. It also illustrates changes in stuttering patterns for the same speaker with different tasks. This comprehensive representation of stuttering frequency and patterns provides a more authentic depiction of stuttered speech for AI models.

*6.1.3* **Emotional Empowerment.** While expert-led data collection were often evaluated and optimized for efficiency [29, 38], StammerTalk data collection was designed and executed with an emphasise on the subjective experiences and emotional empowerment of data contributors. For example, to foster trust and the sense of safety in data contributors, the data collectors - who were also PWS - made the efforts to stutter openly and sometimes voluntarily, during the data collection interviews. The data collectors were also extremely cognizant of the emotional states and stuttering-related struggles of the data contributors, and would swiftly and willingly adapt the interview protocol to accommodate the emotional needs of the data contributors. As evident in the reflections of Rong and Lezhi, both of them were consistently evolving and improving their data collection strategies to provide the participants with a good experience. Without the shared identity and experiences with stuttering, the level of emotional awareness and care demonstrated here would be hard to replicated by data collectors outside the StammerTalk community. In this safe and supportive space created by StammerTalk data collectors, the data contributors were encouraged and liberated to stutter openly, celebrating their stutter as a valuable asset for the dataset, rather than a defect or failure. Such stuttering affirmative attitude has been shown to provide long term emotional and health benefits to people who stutter [35].

As a results, different to the stress, anxiety, and exhaustion often reported in expert-led AI data collection with people with disability [29, 38], data contributors found the StammerTalk data collection sessions enjoyable, relaxing, and empowering. They enjoyed the open conversations with the data collectors, appreciated the empathy and care shown by the data collectors, and often left with greater confidence and self acceptance after the data collection sessions. The data collection process is no longer a transaction between data collectors and data contributors, but a therapeutic and positive experience for both parties.

*6.1.4* **Community.** While expert-led data collection often interact with data contributors individually and separately, community-led data collection drove the community together, building long lasting bonds, connections, and empathy that strengthen the fabrics of the community even after the data collection. As an invisible yet highly stigmatized disability, it is often hard for PWS in China to identify and connect with other PWS in real life [27]. As a result, many data contributors were motivated to participate in StammerTalk's data collection, seeking for a personal connections with the StammerTalk team members. Moreover, as reported in our findings, the connections and conversations with

other PWS empowered the data contributors to see the power of the community and find a sense of belonging and acceptance for their otherwise marginalized identity as PWS. While the interactions between data collectors and data contributors often end with the conclusion of data collection, the relationship between StammerTalk data collectors and contributors tend to last and evolve, as they became more directly connected in the StammerTalk community. As a grassroots, virtual community, such personal ties and connections would be hard to build, but extremely important for the community's long term health and growth.

### 6.2 Urgency to Develop Adequate Socio-technical Infrastructure for Community Data Stewardship

Despite the benefits, there is a significant gap in current socio-technical systems to support initiative like this. We discuss those challenges here.

*6.2.1* **Open-sourcing Datasets.** Open-sourcing datasets has been a significantly more complex process compared to open-sourcing code. This complexity is amplified when the dataset contains sensitive data that cannot be fully anonymized, as is the case with datasets containing highly characterized personal stuttered speech patterns.

The intrinsic value of our dataset for research and AI models lies in the unique speech characteristics of the individuals included. However, the same characteristics that make the data so crucial also render it particularly susceptible to de-anonymization. Unlike other types of data where individuals' features can be blurred or generalized to preserve anonymity, the specific nuances and patterns of speech are themselves the core data points. Removing or altering them would compromise the utility of the dataset.

Furthermore, Personal Identifiable Information (PII) extends beyond just names or addresses. In the realm of speech data, the way someone speaks can, in itself, be a unique identifier. This raises significant ethical and privacy concerns. If malicious actors were to access the dataset, there's potential for misuse or even targeted discrimination against individuals based on their speech patterns. Given these challenges, the responsible handling and potential sharing of such datasets must be approached with caution, taking into account both the scientific value and the ethical implications.

Last but not least, existing legal tools for open-sourcing – such as open-source license – often face limitations when applying to datasets with human subjects and personal information. For example, permissive licenses, such as the Creative Commons (CC) Licenses, has been criticized for the "creative commons loophole" that contributed to the abusive use of personal photographs in training computer vision models [30]. These licenses could also run into conflicts with emerging data and privacy laws that mandates consent from individual data subjects besides data creators [2]. On the other hand, more restrictive licenses, such as Creative Commons non-commercial license, could disincentivize the adoption of the dataset by industry practitioners, thus limit the impact of the dataset.

*6.2.2* **Absence of Legal Framework for Community Data Stewardship.** Traditional personal data protection frameworks are built around distinct roles: data subjects (typically users and consumers), data controllers (often companies), data collectors (platforms or data vendors), and data processors (e.g. annotation and analytical service providers) [42]. These frameworks are constructed on the presumption that each role is performed by separate entities, and legal instruments such as consent forms are formulated to regulate and manage the relationships and obligations between these parties [26].

However, these models fail when the lines blur — when data subjects and controllers are essentially the same, or when the data controller is not a traditional legal entity like a corporation. For instance, StammerTalk, being an unincorporated grassroots community that operates virtually, does not fit neatly into any of these categories. As a result, it is challenging to leverage the default legal instruments - like the participant agreements - to formalize StammerTalk's

data collection activities. Ultimately, a temporary solution was adopted where a few StammerTalk members were designated as data controllers. This is, however, far from an ideal representation of the innate collectivity within the community and creates disproportionate legal liabilities for a few designated members. Additionally, given the fluidity and distributedness of grassroots virtual communities and their membership structure, such an arrangement are prone to break down when members churned or occupied with other activities outside the community.

While there are a few existing proposals for collective data stewardship, such as Data Commons [33] or Indigenous Data Sovereignty [40], those models involves significant political and legal capacities, that are often out of reach for grassroots communities like StammerTalk in practice [42].

*6.2.3* **Navigating Cross-Border, Multinational Personal Data Laws.** Besides the lack of an adequate data stewardship model, the StammerTalk community also needed to navigate the multifaceted web of international data protection laws. Since the StammerTalk community solely exists online — holding meetings via Zoom and group chats, its members are distributed globally across geographical borders. The act of collecting data from community members thus becomes a cross-border undertaking. This results in the need to juggle multiple regulatory frameworks from regions such as the US, EU, and China, each with its nuances, and sometimes, contradictions. The StammerTalk community therefore needed to traverse a labyrinth of legal guidelines, each with its unique stipulations. This complexity not only incurs considerable legal and procedural costs but also poses potential risks. Ensuring compliance with every relevant regulation becomes a daunting task, magnifying the exposure to potential legal liabilities that the community could not afford.

### 6.3 Conditions for Successful Community-led AI Data Collection

A successful community-led AI data collection initiative, like the one demonstrated by StammerTalk, is often influenced by a combination of factors. The question arises: which types of communities are best positioned to embark on similar initiatives? Alternatively, how can we better prepare communities to take on such initiatives? Here we offer some insights based on our case study with the StammerTalk community.

**Technical Expertise Within the Community:** A cornerstone of this project's success was the technical proficiency present within the community. Rong's professional background in speech AI technology endowed him with a thorough understanding of the complexity of the data collection process. His expertise not only influenced the initiative's inception but also ensured that the necessary resources and steps were identified and followed.

**Resourcefulness:** An essential attribute for success is the ability to harness available resources effectively. This initiative was characterized by early partnerships and stakeholder buy-ins, ensuring access to pivotal assets such as annotation services.

**Reputation of Community Organizers:** The standing of the community organizers plays a pivotal role in the project's overall reception and participation rates. When community members trust and respect the organizers, they are more inclined to participate. The positive reputation of the StammerTalk organizers created an environment where members were not only eager to engage but also looked forward to their interactions, keen on acquiring more knowledge and making meaningful contributions.

In summary, the success of such community-driven endeavors is multifaceted, requiring a blend of expertise, resources, and reputation. By maintaining transparency and openness throughout the project, our work aims to further inspire and guide other communities eager to initiate similar ventures.

## 7 LIMITATIONS AND FUTURE WORK

Our work comes with several limitations that require future investigations.

First, **generalizability and scope.** This study revolves around a specific case with a relatively small community leadership. We conducted interviews primarily with two members, which limits the breadth of our insights. While the findings provide valuable insights into StammerTalk, they may not be directly transferable to stuttering communities from other regions or other disability communities at large. Nevertheless, we hope our efforts serve as a catalyst, inspiring other communities to explore this domain with us.

Second, **geographical and language representation.** StammerTalk predominantly represents the Chinese-speaking stuttering community, with the majority of data contributors residing in mainland China and speaking Mandarin Chinese. Other Chinese languages and dialects were not captured in this dataset. Seeing its promise, it would be valuable to generalize this data collection model for stuttered speech datasets in other regions and languages, and understand its efficacy within different cultural and language contexts.

Third, **utilizing the dataset.** To meet the community's expectation, it is urgent and necessary for the AI research community to leverage the StammerTalk dataset to create real change in the experiences of PWS with speech technologies. With StammerTalk community members such as Rong and Lezhi, we plan to first benchmark existing ASR services with this dataset and bootstrap performance improvements through fine-tuning and re-training of state-of-the-art models. Meanwhile, motivated by their desire for more inclusive speech products and services on the market, the StammerTalk community is willing to engage with the broader academia and industry communities in the use and further development of this dataset to catalyze the progress. However, as discussed in previous sections, the concerns remain with the commitment of institutional partners to use the data non-extractively and the ability for the community to effectively exert control and agency over the dataset as well as its derivative products. To address these concerns and facilitate the partnership over - and potential public release of - the community-collected datasets, institutional partners need to take proactive steps to share power and show respect, such as, funding the development of socio-technical-legal infrastructure for community data stewardship, respecting the community's demands for data (co-)ownership and profit sharing, and providing the community with full agency on what and how to collect the data about them.

## 8 CONCLUSION

In conclusion, the rise of AI technologies, while revolutionary, has highlighted glaring disparities in data representation, especially for marginalized social groups such as the disability community. Our research offers an in-depth examination of the grassroots community-led data collection practice using StammerTalk, a grassroots community for Chinese-speaking people who stutter, as a case study. We found that grassroots community initiatives like this is often driven by intrinsic motivations to foster contributions and connections in the community, and can produce AI datasets that authentically represent the community. Community members also gained empowerment, interpersonal skills, and camaraderie from the process, receiving long-term benefits beyond the dataset output. However, challenges arise due to limited resources and the constraints of current socio-technical infrastructures, leading to complexities in navigating international and cross-border data regulations. We thus call for stakeholders – ranging from industries to academia and policymakers – to recognize and invest in building robust infrastructures that empower the disability community in shaping their data practice and data-driven AI experiences.

## REFERENCES


[1] Adriana Alvarado Garcia and Christopher A. Le Dantec. 2018. Quotidian Report: Grassroots Data Practices to Address Public Safety. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 17 (nov 2018), 18 pages. https://doi.org/10.1145/3274286

[2] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical Considerations for Responsible Data Curation. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 55320–55360. https://proceedings.neurips.cc/paper_files/paper/2023/file/ad3ebc951f43d1e9ed20187a7b5bc4ee-Paper-Datasets_and_Benchmarks.pdf

[3] Sam Biddle, Paulo Victor Ribeiro, and Tatiana Dias. 2020. Invisible Censorship. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/. last accessed 2024-01-12.

[4] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[5] Anna Bleakley, Daniel Rough, Abi Roper, Stephen Lindsay, Martin Porcheron, Minha Lee, Stuart Alan Nicholson, Benjamin R Cowan, and Leigh Clark. 2022. Exploring Smart Speaker User Experience for People Who Stammer. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–10.

[6] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective. *ACM Trans. Access. Comput.* 14, 2, Article 7 (jul 2021), 45 pages. https://doi.org/10.1145/3436996

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[8] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[9] James I. Charlton. 1998. *Nothing About Us Without Us: Disability Oppression and Empowerment* (1 ed.). University of California Press. http://www.jstor.org/stable/10.1525/j.ctt1pnqn9

[10] Dasom Choi, Uichin Lee, and Hwajung Hong. 2022. "It's Not Wrong, but I'm Quite Disappointed": Toward an Inclusive Algorithmic Experience for Content Creators with Disabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 593, 19 pages. https://doi.org/10.1145/3491102.3517574

[11] Christopher Constantino, Patrick Campbell, and Sam Simpson. 2022. Stuttering and the social model. *Journal of Communication Disorders* 96 (2022), 106200. https://doi.org/10.1016/j.jcomdis.2022.106200

[12] Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (<conf-loc>, <city>Boston</city>, <state>MA</state>, <country>USA</country>, </conf-loc>) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 8, 13 pages. https://doi.org/10.1145/3617694.3623228

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[14] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I Wouldn't Say Offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. https://doi.org/10.1145/3593013.3593989

[15] Abraham Glasser. 2019. Automatic Speech Recognition Services: Deaf and Hard-of-Hearing Usability. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3308461

[16] Steven M. Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N. Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, Mahima Pushkarna, Joel Riley, Alex Santana, Lei Shi, Rachel Sweeney, Phil Weaver, Ann Yuan, and Meredith Ringel Morris. 2022. LaMPost: Design and Evaluation of an AI-Assisted Email Writing Prototype for Adults with Dyslexia. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 24, 18 pages. https://doi.org/10.1145/3517428.3544819

[17] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.

[18] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331.

[19] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Craig Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of ACL 2020*.

[20] Man Ling Ip, Kenneth O St. Louis, Florence L Myers, and Steve An Xue. 2012. Stuttering attitudes in Hong Kong and adjacent mainland China. *International journal of speech-language pathology* 14, 6 (2012), 543–556.

[21] Rie Kamikubo, Lining Wang, Crystal Marte, Amnah Mahmood, and Hernisa Kacorri. 2022. Data Representativeness in Accessibility Datasets: A Meta-Analysis. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 8, 15 pages. https://doi.org/10.1145/3517428.3544826

[22] Gary Kiger. 1992. Disability simulations: Logical, methodological and ethical issues. *Disability, Handicap & Society* 7, 1 (1992), 71–78.

[23] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2021. Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2986–2999.

[24] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 361, 16 pages. https://doi.org/10.1145/3544548.3581224

[25] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey Bigham. 2021. Sep-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter. https://arxiv.org/pdf/2102.12394.pdf

[26] David Leslie, Mhairi Katell, Michael nad Aitken, Jatinder Singh, Morgan Briggs, Rosamund Powell, Cami Rincon, Antonella Perini, and Smera Jayadeva. 2022. Data Justice in Practice: A Guide for Impacted Communities. https://gpai.ai/projects/data-governance/data-justice-in-practice-a-guide-for-impacted-communities.pdf, note=last accessed 2024-01-15.

[27] Yan Ma, Judith D. Oxley, J. Scott Yaruss, and John A. Tetnowski. 2023. Stuttering experience of people in China: A cross-cultural perspective. *Journal of Fluency Disorders* 77 (2023), 105994. https://doi.org/10.1016/j.jfludis.2023.105994

[28] Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q Nelson, Jordan R. Green, and Katrin Tomanek. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia.

[29] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 52–63. https://doi.org/10.1145/3442188.3445870

[30] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), 1536–1546. https://api.semanticscholar.org/CorpusID:220265500

[31] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[32] Nan Bernstein Ratner and Brian MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders* 56 (2018), 69–80.

[33] Anouk Ruhaak, Greg Bloom, Angie Raymond, Willa Tavernier, Divya Siddarth, Gary Motz, and Melanie Dulong de Rosnay. 2021. A Practical Framework for Applying Ostrom's Principles to Data Commons Governance. https://foundation.mozilla.org/en/blog/a-practical-framework-for-applying-ostroms-principles-to-data-commons-governance/, note=last accessed 2024-01-15.

[34] Arielle M Silverman, Jason D Gwinn, and Leaf Van Boven. 2015. Stumbling in their shoes: Disability simulations reduce judged capabilities of disabled people. *Social Psychological and Personality Science* 6, 4 (2015), 464–471.

[35] Vivian Sisskin. 2023. Disfluency-Affirming Therapy for Young People Who Stutter: Unpacking Ableism in the Therapy Room. *Language, Speech, and Hearing Services in Schools* 54, 1 (2023), 114–119. https://doi.org/10.1044/2022_LSHSS-22-00015 arXiv:https://pubs.asha.org/doi/pdf/10.1044/2022$_L SHSS - 22 - 00015$

[36] Participatory Data Stewardship. 2010. A Framework for Involving People in the Use of Data. Report.(2021, September). Ada Lovelace Institute.

[37] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-First Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. https://doi.org/10.1145/3441852.3471225

[38] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-First Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. https://doi.org/10.1145/3441852.3471225

[39] Seth E Tichenor and J Scott Yaruss. 2021. Variability of stuttering: Behavior and impact. *American Journal of Speech-Language Pathology* 30, 1 (2021), 75–88.

[40] Maggie Walter and Michele Suina. 2019. Indigenous data, indigenous methodologies and indigenous data sovereignty. *International Journal of Social Research Methodology* 22, 3 (2019), 233–243.

[41] Meredith Whittaker, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Elizabeth Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Lisi Rankin, Emily Rogers, Marcel Salas, and Sarah Myers West. 2019. Disability, Bias & AI Report. *AI Now Institute* (20 11 2019).

[42] Janis Wong. 2023. Data Practices and Data Stewardship. *Interactions* 30, 3 (may 2023), 60–63. https://doi.org/10.1145/3589133

[43] Shaomei Wu. 2023. "The World is Designed for Fluent People": Benefits and Challenges of Videoconferencing Technologies for People Who Stutter. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[44] Shaomei Wu, Lindsay Reynolds, Xian Li, and Francisco Guzmán. 2019. Design and Evaluation of a Social Media Writing Support Tool for People with Dyslexia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300746

[45] J Scott Yaruss and Robert W Quesal. 2006. Overall Assessment of the Speaker's Experience of Stuttering (OASES): Documenting multiple outcomes in stuttering treatment. *Journal of fluency disorders* 31, 2 (2006), 90–115.

[46] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[47] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173789