"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*, an online community of Chinese-speaking people who stutter, to autonomously collect and curate stuttered speech data. While marginalized communities were often treated as the *subjects* of personal data collection, StammerTalk pioneered the community data stewardship model to improve their experiences with speech AI technologies. Through interviews and surveys with community members who collected and contributed their speech data, we unpack the unique incentives and challenges marginalized communities face when engaging in grassroots data collection and governance. We find such community-driven data collection not only produced data needed to represent the community in data-intensive applications, but also empowered the community and its members, by enabling authentic expression of their identity through data and strengthening the connections within the community. While current socio-technical systems were not designed to support these initiatives, we discuss strategies and factors for communities to succeed in similar data endeavors.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**.

Additional Key Words and Phrases: AI FATE, datasets, data practice, community data stewardship, 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, 28 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advancements in AI technologies have impacted society and our lives in a significant way. While AI technologies did bring lots of benefits, such benefits have not been shared fairly among different social groups. In fact, increasingly amount of research have identified the potential biases and harms towards marginalized social groups, introduced by AI technologies [25]. As most contemporary AI models were powered by big data, the lack and biased representation of marginalized social groups in AI datasets has been identified as a key driver for AI biases and fairness concerns [6, 26]. An increasing body of research have shown that, when marginalized communities are not adequately included in the data used to train or evaluate AI models, they suffer from the degraded model performance [6, 39] or increased algorithmic risk [14].

This challenge of under-representation in AI data is even more pronounced for the disability community. Not only limited in size and socioeconomic resources, people with disabilities are also often excluded from data collection due to existing physical and digital accessibility issues [26]. Numerous efforts have been made to include people with disabilities in AI data [11, 18, 24, 26, 30]. However, sponsored by tech companies [24] or academic institutions [11, 26, 30], current efforts were mostly designed and led by "experts" outside the disability community, treating members of the

Manuscript submitted to ACM

disabilities as *subjects* and *contributors* rather than the *owner* and *steward* of the collected data. The expert-led data model deprives people with disabilities with their agency and control in the data collection process about them, leads to challenges to engage and incentivize the community to participate [26].

The emerging practice of grassroots data collection presents opportunities for marginalized communities to take initiative and control over their data experiences [1]. In this paper, we present a case study on grassroots data collection initiative led by StammerTalk, an online community for Chinese-speaking people who stutter (PWS). This initiative aimed to create and curate one of the first Chinese stuttered speech datasets to improve their experience with speech AI technologies. By closely following the process from its inception by the StammerTalk community, we collected ethnographic data about this process through observations, interviews, and survey with community members, to understand the process, benefits, and challenges for community data stewardship. Our research questions are:

- RQ1: What are the motivation, experience, and challenges of the community members involved in community-led data collection and how they differ from expert-led practice?
- RQ2: What are the characteristics of the result dataset and how it differs or resembles similar dataset collected by experts?

Our study shows that the community-led data stewardship model not only produce high-quality dataset that authentically represent the community in AI data, but also benefits the community and its members beyond the tangible technical outputs. Contrary to the findings from previous research [26], community members who participated in the community-led data collection were driven by intrinsic goals - such as the making meaningful contribution to the community and connecting with other community members in collective actions, rather than monetary compensation. Consequently, community participants had positive experiences during the data collection, appreciating the empathy, understanding, and personal connections provided by community members who administered the data collection. Beyond the enjoyable experiences, community participants also gained valuable skills, broadened perspectives on stuttering, a sense of empowerment, and stronger bonds that could lead to long-term growth for the community. Our technical evaluation of a subset of the collected data suggests the efficacy of the community-led data collection to produce more diverse and authentic dataset than the traditional approach.

On the other hand, our study also uncovers challenges the community faced, due to limited resources and current socio-technical infrastructure for personal data. Besides the time and energy required for community members to design, execute, quality control the data collection process, they also face structural challenges introduced by existing socio-technical infrastructures that were inadequate for community data stewardship. As a result, the community need to navigate complex geo-political tensions and cross-border data regulations when working with its partners and members across the globe.

Taken together, our work illustrates the benefits and challenges of community data stewardship for AI data, and calls for the attention and investment from industry practitioners, academic researchers, and policy makers to develop the socio-technical infrastructure needed in order for marginalized community to take charge of their data and data-driven experiences.

2 RELATED WORK

2.1 Bias and Under-representation in AI Model Training

A growing body of research showed the under-representation or misrepresentation of marginalized communities in AI training datasets. For instance, Buolamwini and Gebru highlighted the under-representation of women and

darker-skinned individuals in face recognition models [6], a finding later extended by Raji and Buolamwini who audited more commercial systems, revealing widespread skin-type and gender biases across the industry [27]. Similarly, Zhao et al. demonstrated that vision and language models could amplify existing gender biases present in training data if not appropriately constrained [38]. Caliskan et al. further indicated that machine learning algorithms applied to ordinary human language can inherit and perpetuate human-like semantic biases found in web texts, covering a spectrum from morally neutral to problematic biases related to race or gender [7].

In the context of accessibility, people with disabilities have been regularly misrepresented in the data used to train machine learning models [26]. Hutchinson *et al.* revealed a biased representation of people with disabilities in data training text classification models for toxicity prediction that led to over-prediction of disability-related text as toxic [14]. Numerous studies found a degraded performance in computer vision models when used by people with visual impairments as data from this user group is underrepresented in the training and evaluation of these models [18, 39]. Recent work that benchmarked automatic speech recognition (ASR) models identified a significant increase in word error rate (WER) when applied to stuttered speech [3, 20], calling for the inclusion of atypical speech data for training ASR models [21, 24]. Expanding on this concern, Trewin *et al.* argued that AI solutions, especially in critical sectors like employment, education, public safety, and healthcare, possess the potential to either mitigate or exacerbate existing discrimination against people with disabilities [32].

The research emphasized the urgency of considering AI's impact in broader contexts. It is crucial to offer users avenues to rectify errors and ensure the inclusion of marginalized groups, including people with disabilities, in data sourcing and model testing phases. This inclusion is vital for developing truly inclusive and representative AI applications.

2.2 Data Practice in HCI and Accessibility

Having established the challenges of representation in AI training datasets, it's crucial to understand how the field of Human-Computer Interaction (HCI) and accessibility are addressing this by developing rigorous and equitable data collection approaches.

Researchers have increasingly leveraged online crowdwork platforms for gathering data from individuals with disabilities. The rationale behind this approach goes beyond the mere acquisition of data. Such platforms provide people with disabilities with a sense of self-worth, self-efficacy, and autonomy [10]. There's evidence suggesting that participating in these platforms can boost their motivation, both to support scientific research [9] and to derive cognitive benefits from the tasks they engage in [5]. One notable initiative in this direction is VizWiz, which stands out not just for its innovative use of crowdsourcing, but also for its pivotal role in creating datasets originating from people who are blind [2].

Beyond paid platforms, some research efforts rely on volunteer-based data collection platforms, like LabintheWild [22, 23]. Rather than financial compensation, participants are incentivized through personalized feedback. Such feedback serves multiple purposes: from allowing self-comparison with others to facilitating word-of-mouth recruitment as users share their results on social networks. Importantly, online tests have been found to reduce barriers to professional diagnosis for participants, and can help shape their disability identity, fostering self-acceptance. However, a gap remains in terms of two-way communication between participants and researchers, with participants expressing the desire for more interactive feedback and support [22].

Another collective body of work underscores the essential role of accessibility datasets in driving AI applications tailored to the needs of disabled communities. Recognizing the pivotal role of obtaining data from disabled communities, Park et al. delved into designing an online infrastructure tailored to collect AI data from this demographic. Their

study outlined essential guidelines, emphasizing motivation strategies, transparent communication, and accessibility considerations to ensure inclusive AI datasets [26]. Additionally, Kamikubo et al. (2021) offers a broad view, examining the balance of risks and benefits in sharing accessibility data over three decades, highlighting the scarcity of datasets directly sourced from underrepresented communities [16]. Complementing this, both Theodorou et al. and Kamikubo et al. (2023) delve deeper into specific community interactions, the former focusing on blind or low vision participants' contributions for teachable object recognition [31], and the latter exploring the decision-making processes of blind individuals regarding data sharing, factoring in various modalities, objects, environments, demographics and access conditions [17]. Furthermore, Danielle et al. adds a cultural layer to this discourse, addressing the sensitivity of sign language datasets rooted in the unique identity of the Deaf community and the historical implications associated with it [4].

While these studies involve the community in data collection and considerations, none focused on the grassroots aspect our research does – where the community is not just a participant but is intrinsically involved as the steward, initiator, and ultimate decision-maker.

2.3 Grassroots Data Stewardship Models

While considerable research has exlored various aspects of data practices, little attention has been devoted to grassroots data initiatives. grassroots data practices are especially potent in settings where mainstream avenues fail to cater to local necessities. A case in point is the local data practices of Mexico City citizens who harness Facebook for crime reporting and safety information dissemination. Garcia et al. studied these communities and found that administrators play a pivotal role, using crowdsourced data to engage with various stakeholders [1]. Their methods of data gathering, curation, and dissemination amplify trust and foster extended participation, establishing grassroots data as a key tool for tackling unreported safety concerns. This work underscores the role of social media in enabling political action where traditional institutions falter.

However, contemporary data practices and stewardship models are insufficient to provide a comprehensive representation of marginalized communities in AI data. Janis Wong's overview suggests that company-driven data collection and stewardship are riddled with challenges, stemming from a lack of trust, inadequate incentives, and a lack of agency within communities [35]. In the quest for alternative models, the indigenous data sovereignty initiatives undertaken in Australia and Hawaii offer promising insights [34]. Drawing inspiration from such endeavors, our work aspires to build a community-led data model, specifically catering to the stuttering community.

While the participatory data stewardship model stands as a benchmark in the realm of data collection and management [29] [1], our methodology endeavors to transcend its boundaries. In our paradigm, the community transcends the traditional role of mere participants; they emerge as the initiators and leaders, emphasizing a profound shift in data stewardship dynamics.

2.4 Stuttering and Speech AI

Individuals who stutter often confront evident performance discrepancies when interacting with current ASR-enabled applications. While recent technical advancements have shown potential in enhancing ASR models for stuttered speech, the scarcity of stuttered speech data remains a significant barrier to progress.

[1] https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

2.4.1 **ASR's Inadequacy for Individuals Who Stutter.** Bleakley et al.'s recent user experience research illuminates the pronounced challenges faced by people who stutter when interacting with smart speakers [3]. Not only do these individuals encounter inaccurate transcriptions, but ASR systems often interrupt them mid-speech or, at times, completely fail to respond to their requests. Such behavior is not merely a technical glitch; it effectively imposes structural barriers that restrict access to essential services and information for those who stutter. Echoing this sentiment, Clark et al. emphasize the necessity of accommodating diverse speech patterns in ASR development and outline the fundamental challenges tied to creating truly inclusive speech interfaces [8].

Highlighting the urgency of these findings, Apple's research team, led by Lea *et al.*, recently conducted benchmarks on their ASR systems specifically for stuttered speech. Their results unveiled a stark performance discrepancy, with a tenfold increase in Word Error Rate (WER) for users with severe stuttering. Moreover, when evaluating the performance of endpointer models in terms of truncation rate, the disparity remains evident: a significant proportion of utterances from people who stutter are prematurely cut off compared to those from the general population [20]. This finding is consistent with previous observations by Lea *et al.*, further emphasizing the pressing need for advancements in this domain [21].

2.4.2 **Existing technical efforts.** The endeavor to enhance ASRs for stuttered speech has been primarily characterized by two main thrusts: i) the establishment and curation of stuttered speech datasets, and ii) the modification of pre-existing ASR models to improve their proficiency in detecting and understanding stuttered speech.

i) Datasets of Stuttered Speech. With some originally conceived as tools for speech therapy, several datasets featuring speech samples from individuals with stutters have been compiled. These include the FluencyBank [28], the University College London's Archive of Stuttered Speech (UCLASS)[12], the SEP-28K dataset sourced from public podcasts[21], non-standard speech from Project Euphonia [24], and the LibriStutter dataset that comprises synthesized stutters [19]. Despite the substantial size of these datasets, they contain only a limited portion of stuttered speech samples. Furthermore, inconsistencies in stutter-specific annotations present challenges, both of which hinder the effective training of advanced ASR models for stuttered speech.

ii) Adapting Existing ASR Models for Atypical Speech. Google has been at the forefront of these efforts via Project Euphonia [24], making significant strides in refining their existing model using a relatively modest volume of atypical speech data. However, their research has predominantly concentrated on recognizing brief phrases (such as command words) and non-stuttered speech forms, such as speech patterns of people with ALS or Down Syndrome. As a result, the transferability of their advances to contexts involving conversational and/or stuttered speech remains uncertain. In addition to releasing the Sep-28K dataset [21], Apple has also explored the tuning of ASR models for dysfluent speech. Lea *et al.* have recently compared three interventions to augment ASR systems, some of which require minimal stuttered speech samples [20].

In summary, despite the increasing interest from major technology companies in addressing this challenge, progress has been gradual, and the existing gap remains. The challenges faced by individuals who stutter are manifold, encompassing both personal and systemic dimensions. It's imperative to take actions to ensure AI systems don't further exacerbate disparities or suppress certain demographics.

3 BACKGROUND

Here, we provide an overview of the StammerTalk community and its members to contextualize this project, detailing the data collection process and steps undertaken. The information was sourced from public channels, such as StammerTalk's WeChat profile as well as interviews with key community members.

3.1 Community Background

3.1.1 **Introduction and Platforms.** StammerTalk (口吃说) is an online community serving Chinese-speaking people who stutter. Originating as a podcast, it featured interviews of individuals with stutters by fellow stutterers. The community also manages a public WeChat[2] official account, sharing personal stories and scientific insights on stuttering, raising awareness among the Chinese stuttering community. In addition, StammerTalk moderates a WeChat group dedicated to individuals with stutters, with approximately 500 members primarily from mainland China, serving as a vital platform for these individuals to connect.

3.1.2 **Activities and Events.** In addition to its regular initiatives, StammerTalk hosts bi-weekly online self-help groups since 2020. The community also organizes occasional events, such as a virtual conference in October 2022, celebrating the International Stuttering Awareness Day with research presentations, round table discussions, and open mics.

3.1.3 **Organizational Structure.** Functioning without legal registration in any country, StammerTalk operates through volunteers. A dedicated team of 10 volunteers, split between China and overseas, shares the operational responsibilities. These range from hosting meetings, podcast production, blog writing, group moderation, to event management. Though the community lacks a formal structure, tasks are divided based on each volunteer's interest, skill set, and availability. For instance, a volunteer with expertise in Speech-Language Pathology leads the self-help group sessions. Three of the 10 volunteers, forming a "core team", are responsible for new community initiatives and community expansion. With most communication and coordination taking place in WeChat, they meet bi-monthly via video conferencing to assess ongoing ventures and strategizing for the future.

3.1.4 **Funding and Membership.** StammerTalk operates on a modest budget, drawing funds from tips on its WeChat public account and personal donations from volunteers. No volunteer receives compensation for their efforts.

Similar to other online communities, StammerTalk's membership remains fluid, with engagement spanning various platforms such as the podcast, WeChat group, and occasional events. For this study, we've focused on members of the StammerTalk WeChat group and followers of their public account, recognizing these are the two major channels where the StammerTalk volunteers interact with and mobilize its members.

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 [15]. 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 Project Background

Here we describe how the data collection project was conceived and executed by the StammerTalk community.

[2]WeChat (微信) is a messaging and social media app developed by Tencent, popular among Chinese speakers globally.

3.2.1 **Conception.** In December 2022, the concept of gathering a Chinese language stuttered speech dataset emerged during a discussion in a WeChat group chat between the StammerTalk leadership and one author of this paper. The primary motivation was to utilize this dataset as benchmarking resources to enhance ASR models' performance for stuttered speech. While the proposal was met with enthusiasm and appreciation from all group members, Rong, a key member of the StammerTalk leadership, graciously stepped forward to spearhead the project's planning and execution.

3.2.2 **Preparation.** To set the groundwork for data collection, the StammerTalk core team undertook several pivotal steps:

(1) **Build partnerships.** From December 2022 to January 2023, StammerTalk identified essential resources and established partnerships with academic, nonprofit, and corporate entities specializing in AI technology, stutter support, and socio-technical frameworks.
(2) **Secure Annotation Support.** Leveraging a connection from a renowned academic institution in China, Rong successfully engaged a commercial speech annotation company, securing their commitment to voluntarily annotate the collected data.
(3) **Develop Annotation Guidelines.** Given the absence of guidelines for annotating stuttered speech in Chinese, Rong collaborated with the aforementioned company to draft them, and to provide training for non-stuttering annotators. Starting with the company's existing guidelines, they integrated stutter-specific instructions from English stuttered speech literature [21]. The guideline was finalized after three refinement iterations based on various feedback over a month.
(4) **Procure Legal Assistance.** Ensuring the initiative's legal compliance, StammerTalk sought pro bono legal expertise via a partnership with a US-based nonprofit ([name redacted]). This team provided guidance on participant agreements, data privacy, security, and regulatory concerns.

3.2.3 **Participant recruitment.** Meanwhile, StammerTalk began its participant recruitment for the data collection through a detailed post on its WeChat official account. The main contents of this recruitment post included:

(1) **Objective.** The post highlighted the challenges PWS (People Who Stutter) often face with speech AI systems and motivated community members to contribute their speech data to enhance these systems' understanding of stuttered speech.
(2) **Eligibility.** The only criterion was being a person who stutters, with no other restrictions such as age or gender.
(3) **Participation Process.** Potential participants were directed to scan a QR code to connect with the StammerTalk team. They were informed that a data collection session would approximately last 90 minutes.
(4) **Recording Specifications.** The post clarified that the session would be virtual, either via Zoom or Tencent Meeting. Participants were advised to use a quiet location and an external microphone for optimal audio quality.
(5) **Data Privacy.** Assurances were given that any identifiable information would be removed from recordings, and the data would be open-sourced solely for non-commercial purposes.
(6) **Compensation.** Participants were informed of a ¥100 RMB ($14 USD) cash incentive delivered via WeChat and a commemorative backpack sponsored by the commercial data annotation firm.

The initial recruitment effort attracted over 40 community members. A subsequent recruitment phase, using the same strategy, was initiated in July 2023.

3.2.4 **Data collection.** Upon signing up for a data collection session, community members were presented with a participant agreement form. This form outlined 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 acknowledged and signed, participants were scheduled for a structured 90-minute data collection session conducted via Zoom or Tencent Meet, broken down as follows:

(1) **Preparation (30 mins)**: The session commenced with an introduction from the StammerTalk core team member (the interviewer) hosting the meeting. The interviewer then briefed the participant (the interviewee) on the session's structure. Additionally, the interviewer ensured optimal technical and environmental conditions for the recording. Participants were then requested to complete the Overall Assessment of the Speaker's Experience of Stuttering (OASES) [37], the results of which were paired with the recordings as metadata.
(2) **Unscripted Spontaneous Conversation (30 mins)**: A casual conversation took place, focusing on the participant's background and personal stuttering experiences.
(3) **Voice Command Recitation (30 mins)**: Participants were provided a set of typical voice commands to read aloud.

The latter two components of the session were audio recorded. Subsequently, these recordings were securely stored in a shared Google Drive folder, accessible only to the StammerTalk core team and selected partners. Approximately an hour of speech data was collected from each session. Trained annotators executed the speech-to-text transcription and stuttering annotations in 10-hour batch intervals.

By August 2023, a total of 58 StammerTalk community members had participated in this data collection process. For the rest of this paper, we will refer to the StammerTalk core team members who collected and processed the data as **data collectors**; the community members who participated in the recording sessions are denoted as **data contributors**.

3.3 Positionality Statement

Recognizing that as researchers, our personal backgrounds and identities shape how we engage with communities 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, both directly through company's or lab's platforms, and 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 ties with the StammerTalk moderators and collaborated on other stuttering-related projects beyond this data collection endeavor.

Though our close relationship with StammerTalk and shared experiences as stutterers fostered 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 METHOD

4.1 Semi-structure 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.

Table 1. **Background Information of Data Collectors**

Name	Gender	Age	Country	Occupation	Community Role
Rong	M	25-35	Austria	Research scientist in a speech technology company	StammerTalk co-founder, core team member
Lezhi	F	25-35	US	Data scientist in a large retailer company	StammerTalk early member, core team member

Interview Setup. 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 are shared in Table 1. Per the preference of the data collectors, we will refer them with their real names.

Interview protocol. The interview initiated with a **warm-up** session where the data collectors share about their professional roles and describe their personal experiences and challenges related to stuttering. Then we delved into the **motivation and incentives**. The data collectors were asked about their inspiration or driving force behind participating in this initiative. This was followed by a detailed exploration of the **processes and experiences**. The data collectors provided insights into their preparation, planning stages, and the execution of various tasks. They discussed the distribution of responsibilities, technical setup, participant recruitment strategies, anticipated workloads, and timelines, among other topics. To better understand the nuances of the endeavor, the data collectors were also questioned about any deviations from the initial plan, unforeseen circumstances, or lessons gleaned from the overall process. The interview then transitioned to highlight the **challenges and strategies**. Moderators were encouraged to reflect on both anticipated and unexpected hurdles and the strategies they employed to overcome them. Finally, the interview wrapped up by prompting the data collectors to **introspect** on their journey, emphasizing lessons learned, personal growth, and potential future plans. They were given an open platform to share any additional insights or anecdotes not covered earlier, ensuring a comprehensive understanding of their experiences.

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.

4.2 Survey with data contributors

Our initial interviews with the data collectors provided detailed insights into the data collection processes, and the inherent challenges and dynamics of moderating interviews with people who stutter. These narratives significantly enhanced our understanding, addressing our preliminary research questions. However, solely relying on the perspectives of the data collectors might leave out the rich experiential insights of the data contributors themselves – those who stutter and are willing to contribute their data. Recognizing the potential depth that these data contributors could offer

and to ensure we received a holistic view, we designed a survey for these individuals. This methodological expansion aimed to validate and potentially enrich the themes that emerged from the moderator interviews, ensuring their relevance and resonance with the larger data contributors pool. By consolidate perspectives from both data collectors and data contributors, we sought to capture a comprehensive landscape of the data collection process and its nuances. The survey was conducted in Mandarin, and the results presented in subsequent sections are translated into English.

Survey Questions. The survey design was informed and shaped by insights garnered from prior interviews with data collectors. The survey comprised 14 distinct questions, both open- and closed-ended. For a comprehensive view of the entire survey, please refer to the Supplementary Material. The survey can be 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.

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 Tecent 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 [13]. 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 qualitative 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 [26], 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 aimed to use their skills to address the community's technological challenges. Rong, associated with a commercial speech technology company and leading this project, emphasized:

> *I'm professionally involved in this space, understanding the entire process well... The various stages like data collection, annotation, model training, and evaluation are part of my daily jobs. Therefore, undertaking this project end-to-end would be intriguing and meaning for me.* (Rong)

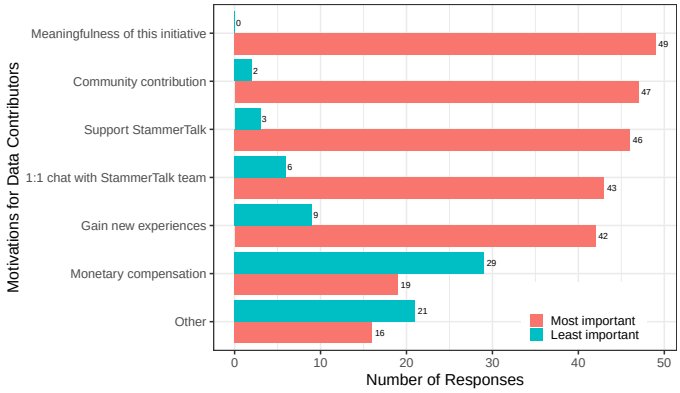

Fig. 1. The most and least important reasons for data contributors to participate in data collection project.

Upon realizing the gap in Chinese stuttered speech data, both Rong and Lezhi recognized the dataset's potential for research, education, and technical work around stuttering, benefiting the stuttering community in China. 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.* "

Both Rong and Lezhi also perceived the project as enriching their professional journey. Rong, primarily involved in speech technology R&D, viewed the end-to-end experience, starting from data collection, as valuable for his professional growth. Meanwhile, Lezhi believed that showcasing a project at the intersection of technology and stuttering would help her transparently communicate her stuttering to peers and potential employers. Both Rong and Lezhi envisioned the data collection project as an act of self- and community advocacy. As Lezhi elaborated on her perspective:

> *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. (...) (Compared to people who do not stutter,) 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)

Similar to the data collectors, most of the StammerTalk community members who participated in the project as data contributors were not driven by material gains or external recognition, but their recognition of the value of this project to the stuttering community and their desire to contribute to and connect with the community. As observed by Rong and Lezhi,

> *Most of them (data contributors) found this project very meaningful. Even though a lot of them did not fully understand what speech AI is, they know that this project is going in the right direction that will help the entire stuttering community and the research around stuttering.(...) A very small percentage of people were motivated by the monetary rewards.* (Rong)

> *They (data contributors) came with questions. (...) They wanted to know more about our project, about StammerTalk, and about my personal experience with stuttering.(...) Many people do not have someone in their lives to talk to about stuttering. So they considered me as a listener, or even as a mentor, a mentor for stuttering.* (Lezhi)

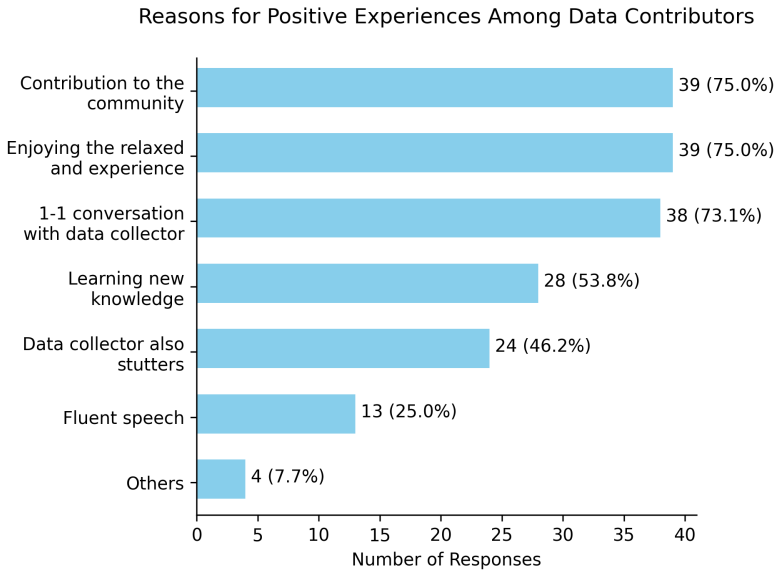

Fig. 2. The primary reasons that led to the positive experiences among the data contributors in the data collection project.

These observations were confirmed by the data from the contributor survey. As shown in Fig. 1, when asked to pick the most and the least important reasons for them to participate in this project, more than 80% of the 55 survey respondents found their top motivators to be: their recognition on the value of this project ("*meaningfulness of this initiative*", N=49), contributions to the stuttering community ("*community contribution*", N=47), support StammerTalk's projects ("*support StammerTalk*", N=46), opportunity to talk to StammerTalk team one-on-one ("*1:1 with StammerTalk team*", N=43), and the expectation to gain new and interesting experiences ("*Gain new experiences*", N=42). A relative small number (N=19/55) of the survey respondents considered "*Monetary compensation*" as the most important reasons for them to participate. In fact, "*Monetary compensation*" was the most frequently picked (N=29/55) as the least important reason(s) to participate in the data collection.

To sum, the StammerTalk community demonstrated a great amount of agency and autonomy to initiate and participate in the data collection project. Leveraging the existing technical talents and collaborative structure within the community, community members led and participated in the data collection to make a meaningful contribution to the community, advocate for their needs and rights, build deeper connection with each other, and embrace their identity as people who stutter.

5.2 Experiences

5.2.1 ***Overall Experience.*** Despite research suggesting that individuals who stutter often experience heightened stress during speech-related tasks [36], the majority of StammerTalk community members found their 1.5-hour recording sessions both positive and enjoyable.

52 out of the 55 (95%) 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 highlight the primary factors

contributing to their feelings. The summarized responses can be found 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, highlighting the value of community engagement and fostering deeper connections among its members for the data contributors.

5.2.2 ***Positive Experiences of Data Contributors - Role of Empathetic Data Collectors.*** Most data contributors found their interaction with data collectors during the data collection process uniquely positive, greatly contrasting with their typical speaking experiences. The ambiance of trust and comfort during the interviews largely stemmed from the empathetic nature of the data collectors.

A significant number of data contributors (N=23/54) appreciated being interviewed by someone who also stutters. As Rong observed, the mutual experience of stuttering established an immediate sense of trust. The participants often remarked, " *oh, you also stutter!', followed by, 'now I can relax.'* " Lezhi had similar observations, " *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.* "

Beyond this shared identity, the data collectors took deliberate steps to nurture a positive environment. They shared personal experiences, adapted conversation topics to accommodate interviewee's mental and emotional state, and avoided any behavior that might be perceived as intimidating. Lezhi remarked on the importance of this approach, emphasizing her ability to discern when a participant might be feeling nervous and adjusting her approach 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)

Survey results reaffirmed that the data collectors have successfully connected with, emphasized, and accommodate the data contributors during and after the interview. 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 (N=51/54), clear communication about the data collection process (N=45/54), and the substantial empathy shown by the interviewers (N=38/54).

Interestingly, the supportive environment meant some data contributors spoke more fluently than usual during the recording session, with N=13 out of 54 citing this increased fluency as a factor in their positive experience. While all participants identified as people who stutter, some were still self-conscious about it and naturally aimed for fluency. Such a focus on fluency might be attributed to the physiological and emotional challenges associated with stuttering [36]. However, this increased fluency could result in the collected data being less representative of typical stuttering patterns. This potential limitation will be explored further in the "Challenges" section below.

5.3 Gains

Beyond the direct benefit of curating an immediate, tangible data asset for the community, participants of this project also gained valuable skills and experiences, broadened perspectives, and connections that could lead to long-term positive outcomes for the community.

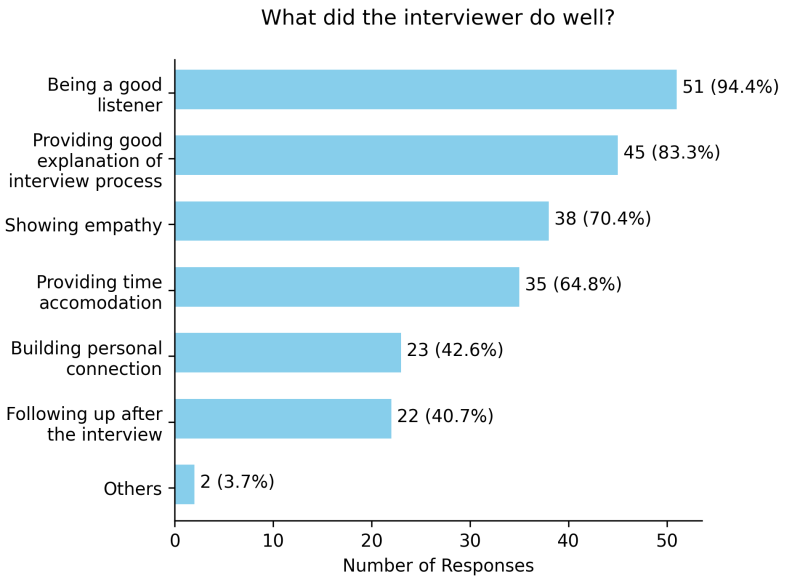

Fig. 3. Data contributors' feedback on data collectors' competencies during the data collection project.

Data Collectors' Insights. Although neither Rong nor Lezhi sought monetary rewards from the project (with Rong even making a personal financial contribution), they cited personal and professional growth in several areas: 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 hear 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 weren't transient. Both Rong and Lezhi continued to foster personal connections with many contributors post-interviews through platforms like WeChat.

Perspectives of data contributors. In our analysis, despite receiving a modest monetary compensation ($14 USD) for their contributions, none of the data contributors cited this as their primary gain from the project. Instead, their most significant takeaways were threefold: 1) a sense of unity, recognition, and empowerment within the stuttering community; 2) the profound emotional relief, self-acceptance, and personal growth from genuine, empathetic dialogues; and 3) enhanced understanding about stuttering. These benefits align with, and even surpass, their initial motivations for participation."

Some participants (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 progress. One participant 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: " *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.* " Collectively, these reflections emphasize the transformative potential of community-driven initiatives in fostering a deeper sense of belonging and empowerment within the stuttering community as well as within broader societal frameworks.

Participants (N=14) also gained immensely from the genuine, one-on-one communication opportunities provided to them. Their feedback consistently underscores the immense personal growth, realization of their inherent potential, and emotional relief they experienced by having a space where they could " *communicate with others and open one's heart* " and " *have the courage to face the real self and accept one's stuttering behavior.* " 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 fueled their optimism. Through these interactions, it becomes clear that participants received not just an avenue to express, but also deeper self-awareness, acceptance, and a renewed sense of empowerment — driven by the profound empathy and connection they experienced.

Other participants (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 – it effectively conveyed their message to me. This deeply resonated with me, and I am currently learning this way of speaking.* "; similarly, P8 mentioned: " *The biggest takeaway was being interviewed by someone who stutters like me and conversing with them. They briefly shared some methods on how to alleviate awkwardness and how they cope with stuttering, as well as how to relieve anxiety. Through our conversation, I learned quite a lot.* "

In summary, the data contributors greatly valued their participation in the data collection project for the interpersonal connections, empowerment, and advocacy it fostered. Similarly, data collectors experienced personal growth and formed lasting connections, emphasizing the project's profound impact beyond its primary objective.

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.*** There were four major challenges data collectors had to navigate through.

(1) Time Commitment. First, the time commitment has been one of the biggest challenges to make progress at a desired pace. Rong and Lezhi, both employed full-time, could only allocate evenings and weekends to the project, affecting their personal and family time. Given their locations in Europe and the US, coupled with the primary participant base in China, scheduling was particularly challenging due to time zone differences. Their limited availability and the often last-minute rescheduling requests resulted in, at most, one or two recording sessions per week. Given the significant amount of time and energy required to recruit and schedule for the data contributors, and the fact that " *for the second and the third batch, there would be probably fewer people signing up,* " Rong anticipated it might push the project's completion timeline to potentially over a year to reach the goal of 100 hours from 100 individuals.

(2) Data annotation. Second, 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. It took three iterations for the annotators to be able to identify and label the stuttering events. During each iteration, Rong would carefully review the annotations produced by the annotators, and returned with corrections with detailed explanations. While 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 strived to have the data sufficiently cover **the variety in stuttering patterns and severity levels**. Stuttering, being multifaceted, varies in frequency, severity, and manifestation across individuals and contexts [20, 36]. 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 wouldn't, 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. Contributors weren't pre-screened for speech samples or stuttering descriptions. While they did complete the Overall Assessment of the Speaker's Experience of Stuttering (OASES) [37], it wasn't a selection criterion but rather treated as metadata. Rong reflected upon the selection criteria, and emphasized that a person's self-identification as someone who stutters should be the sole requirement for participation, making the sample intrinsically representative of stuttering community and eliminating potential biases. This approach accentuates the difference between community-led and third-party 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 identified their stuttering during the interview as a challenge they faced. Another 13 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 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. Some 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.

In summary, we identified key obstacles in community-led data collection: substantial time commitments, the necessity of resources for speech data annotation, legal and privacy uncertainties, and finding the balance between stutter data representativeness and data contributors' comfortness.

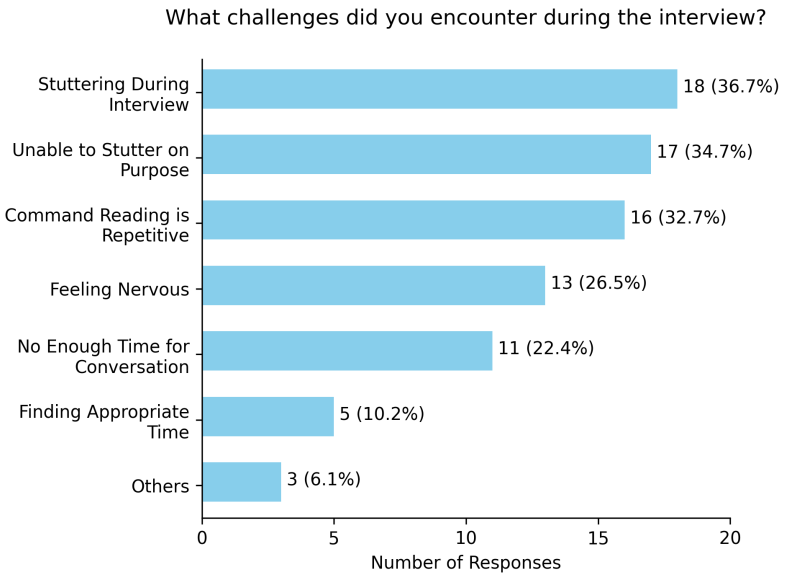

Fig. 4. Primary challenges faced by data collectors during the data collection process.

6 TECHNICAL EVALUATION OF OUR INITIAL DATASET

As of August 2023, the StammerTalk team shared with us the first batch of speech recordings and annotations of 10 participants. Here, we provided a summary statistics of our dataset, and compared our dataset with previous stutter-focused dataset, such as SEP-28k [21], to show the validity and higher diversity of speech and stuttering events our dataset captured.

6.1 Annotation

As described in **Project Background** (Section 3), the StammerTalk together with the commercial speech annotation company developed a detailed annotation guideline tailored for Mandarin speech to identify and categorize stuttering events. This guideline was informed by the annotation guidelines for creating SEP-28k [21], and was established in Chinese to ensure clarity and precision for our native Chinese annotators. The annotation categories were as follows:

- **[]**: **Word-level repetition**. Designated for marking entire repeated words or phrases. If a repetition only concerns a singular phoneme, the tag **/r** is applied.
- **/b**: **blocks**. Used to annotate evident prolonged blocks or brief interruptions.
- **/p**: **prolongation**. Denotes prolonged phonemes.
- **/r**: **sound repetition**. Intended for the annotation of repeated sounds, such as a lone consonant or vowel, that do not constitute an entire word.
- **/i**: **interjections**. Marks unnatural utterances like '嗯', '啊', or '呃'. Notably, naturally occurring interjections that don't disrupt the speech flow are excluded from this annotation.

This structure was designed to provide a nuanced understanding of stuttering patterns in the Mandarin-speaking population and facilitate accurate and consistent annotations throughout our dataset.

Table 2. **A summary statistics of stuttering patterns in voice command reading dataset, and the distribution of stuttering types across 10 participants. It showed that stuttering patterns are unique across individuals (lowest and highest value of each metrics highlighted in the table).**

PID	% of Stuttered Commands	Stuttering Rate per Min	Stuttering Rate per Command	[]	/b	/p	/r	/i
D1	33%	3.05	1.77	17%	17%	11%	53%	2%
D2	22%	3.74	1.22	80%	5%	0	15%	0
D3	10%	1.19	1.72	10%	56%	0	34%	0
D4	2.0%	0.41	1.41	35%	35%	0	24%	6%
D5	31%	5.49	1.38	11%	9%	5%	8%	68%
D6	97%	9.18	3.40	55%	27%	6%	9%	2%
D7	46%	5.09	1.10	99%	1.0%	0	1.0%	0
D8	25%	3.14	1.62	55%	32%	1%	11%	2%
D9	60%	8.54	1.80	49%	27%	12%	13%	0
D10	99%	7.24	5.65	51%	20%	13%	16%	0
Aggregate	**38%**	**4.62**	**2.76**	**51%**	**22%**	**8%**	**15%**	**4%**

6.2 Voice Command Reading

We first investigated the sub-dataset comprising stuttered speech recordings where each participant was tasked with reading a set list of voice commands. The total number of commands read by each participant ranged from 208 to 523, with a median count of 283. Participants spoke for 24 minutes on average (SD=5.8 min), ranged from 15.3 to 31.3 minutes. Note that the variation of the total number of commands for data contributors comes from the fact that they may be asked the repeat the same command multiple times and voluntarily stutter. We have discussed the reason behind this design choices and will discuss its implication at the end of the section.

In terms of stuttering events, the percentage of stuttered commands (calculated by *# of stuttered commands* / *# of total commands*) varied with a low of 2% and a high of 99%, averaging at 42.5% (SD=33.5%), emphasizing the variability in stuttering frequency/severity across our participants. Furthermore, the stuttering rate was calculated by dividing the number of stuttered events by the speech duration, revealing a mean rate of 4.7 stuttering events per minute (min=0.4, max=9.18, SD=2.96). Last but not least, because there can be multiple stuttering events in each command reading, we showed that the average stuttering events per command was 2.17 (SD=1.34), ranging from 1.11 to 5.65 stuttering events per command across participants. A detailed per-participant statistics is presented in Table 2. In the table, we further broke down stuttering types and showed their distribution: we found that stuttering behaviors display notable variability in their patterns across individuals. For instance, while participant D2 predominantly exhibited word-level repetitions, accounting for 80% of their stuttering events, participant D5 demonstrated a markedly different pattern, with 68% of their stuttering events being interjections (filler words). This variability emphasizes the uniqueness of stuttering patterns across individuals when reading commands.

6.3 Unscripted Spontaneous Conversation

The second part of our dataset consists of around 30 minutes of free conversations between the data collector and the data contributor. We transcribed and annotated the entire conversation for both data collectors and data contributors, but only reported on data contributors' speech annotation here for consistency. They spoke for 16.57 minutes on

average (SD=4.35), with a min of 10.12 minutes to a max of 25.33 minutes. Data annotators used Praat [3] to annotate the conversation speech and the data was segmented in arbitrary lengths (sentences, phrases, etc.) by subjective natural poses. Thus, we calculated the number of total segments (mean=125.5, SD=29.7, min=78, max=195) and the total number of stuttered segments (mean = 97.60, SD= 20.22, mean=65, max=122).

To show the diversity of stuttering patterns as well in our free conversation dataset, we presented the proportion of stuttered segments, stuttering rate per min, and stuttering rate per segment, along with distribution of all stuttering types for each individual in Table 3. Similar to command recitation, we observed a high variation in each of these metrics.

Table 3. **Summary of stuttering metrics in the unscripted spontaneous conversation across 10 participants, with the lowest and the highest value of each metrics highlighted.**

PID	% of Stuttered Segments	Stuttering Rate per Min	Stuttering Rate per Segment	[]	/b	/p	/r	/i
D1	82%	13.07	2.78	57%	12%	4%	10%	18%
D2	70%	15.85	2.04	29%	1%	10%	19%	41%
D3	58%	14.29	1.41	53%	10%	4%	2%	30%
D4	92%	36.55	4.82	43%	26%	1%	15%	16%
D5	70%	14.22	1.66	85%	6%	0%	3%	5%
D6	97%	34.89	3.55	42%	6%	14%	8%	30%
D7	83%	18.37	2.38	68%	11%	1%	4%	16%
D8	77%	13.07	1.96	90%	1%	0%	5%	4%
D9	62%	8.92	1.23	52%	14%	3%	3%	28%
D10	99%	46.62	6.21	39%	10%	12%	5%	33%
Aggregate	**78%**	**20.69**	**2.73**	**46%**	**22%**	**2%**	**19%**	**11%**

6.4 Comparison with Previous Dataset

A handful of stutter speech datasets exist in the literature, such as FluencyBank [28], Project Euphonia [24], University College London's Archive of Stuttered Speech (UCLASS)[12], and the synthesized stutter dataset LibriStutter[19]. However, these datasets often face limitations due to inconsistent annotations or even a complete absence of them. A contemporary and more relevant dataset for our work is Stuttering Events in Podcasts (SEP-28k) [21]. Notably, it contains over 28,000 clips labeled with five event types matching our annotations. In this section, we juxtapose our dataset with SEP-28k to underline its validity, comparability, and potential advantages.

A primary distinction between the two datasets is the nature of the annotations. SEP-28k adopts time-interval-based annotations, whereas our approach is event-based [33]. This methodological disparity necessitated the development of specific metrics to facilitate a meaningful comparison:

- **Stuttering rate per minute**: Computed by dividing the total number of stuttering events by the speech duration.
- **Distribution of Different Stutter Types**: This metric illustrates the percentage distribution of each stuttering event type, determined by counts of specific stutter type/total stuttering events $\times 100$. A significant distinction

[3]https://www.fon.hum.uva.nl/praat/, a computer software package for speech analysis in phonetics.

Table 4. **Comparison between our dataset and SEP-28k [21] by stuttering rate per minute and by distribution of event types among all stuttering events.**

	Stuttering rate per minute	Event Type Distribution				
		[]	**/b**	**/p**	**/r**	**/i**
SEP-28k	2.4	23%	28%	23%	19%	49%
Conversation	20.69	46%	22%	2%	19%	11%
Voice Command Reading	4.62	51%	22%	8%	15%	4%

between our dataset and SEP-28k pertains to segmentation: While SEP-28k utilizes 3-second audio clips, our dataset segments speech based on natural occurrences—either by a distinct command or by a naturally occurring pause in conversation.

Further distinguishing our work from SEP-28k is the language focus. Our dataset pertains to the Chinese language, which might intrinsically carry differences in pronunciations, speech patterns, and the nature of stuttering, given its tonal nature and unique linguistic structures. This language difference introduces an added dimension to the comparison, underscoring potential linguistic and cultural variations in stuttering manifestations.

Results are presented in Table 4. In comparing our dataset with the SEP-28k dataset, there are both resemblances and marked contrasts in terms of the stuttering rate per minute and the distribution of event types. Specifically, the SEP-28k dataset exhibits a stuttering rate of 2.4 events per minute, which is considerably lower than the rate noted in our Conversation subset at 20.69 events per minute. Our Voice Command Reading subset, while higher than SEP-28k, records a more moderate rate of 4.62 events per minute. This disparity can be attributed to the nature of the source data; the SEP-28k dataset is derived from podcasts, many of which discuss speech disorders (from both people who stutter or do not), but may not necessarily contain high rates of actual stuttered speech [21]. Our datasets, both the Conversation and Voice Command Reading subsets, are designed to encapsulate more genuine stuttering occurrences, with the latter even involving instances where participants were sometimes asked to voluntarily stutter. This design choice can explain the pronounced difference in rates.

In the realm of stuttering type distribution, both datasets exhibit distinct and similar trends. Word-level repetitions, symbolized by **[]**, are prominent in both datasets. Similarly, the distributions for **/b** blocks and **/r** sound repetition are relatively consistent across datasets, with marginal variations. The presence of **/p** prolongation in our dataset is considerably lower (2% and 8%) compared to that of SEP-28k, possibly due to the fact that Mandarin, being a tonal language, and English possess distinct phonetic and phonological properties. One of the most striking differences emerges in the interjection (**/i**) category. SEP-28k has an elevated interjection rate at 49%, a figure that stands in stark contrast to our dataset's 11% for Conversation and 4% for Voice Command Reading. A plausible reason for this disparity can be traced back to the differential annotation guidelines or definitions for interjections between the datasets. While SEP-28k classifies interjections as filler words like "um," "uh," and "you know", our guidelines mark them as unnatural utterances – the guideline specifically excludes naturally occurring interjections that seamlessly blend into the speech flow without causing disruptions. Such distinctions and commonalities in event type distribution not only reflect the innate linguistic and cultural contrasts between English (SEP-28k) and Mandarin (our dataset) but also the nuances brought about by divergent annotation strategies and definitions.

6.5 Discussion and Summary

Collecting speech data with diverse stutter patterns is paramount for understanding and addressing stuttering in speech systems. Our analysis confirmed this, revealing high variances in stutter patterns across participants. This validates our approach to collate data from a larger cohort of 100 individuals rather than amassing 10 hours of data from just 10 participants. Feedback from data collectors resonates with this sentiment. As Rong noted, achieving the target might require more than a year given the constraints, emphasizing a measured, consistent pace of one or two recording sessions weekly. However, despite these challenges, the commitment is crucial and justified. A broad participant base ensures diverse speech samples, which is essential for comprehensive analysis and system training.

When we delve into the individual data, disparities emerge in stutter patterns between Voice Command Reading tasks and Natural Conversations, even within an individual participant. This distinction accentuates the importance of capturing both datasets. The Voice Command Reading tasks, although inclusive of voluntary stutters, guarantees the dataset harbors ample events and diversity, vital for effectively training AI speech models. In contrast, the unscripted conversations provide a more organic glimpse into stuttering patterns, thereby enhancing our understanding of the phenomenon.

Our dataset's comparison with pre-existing datasets showed both similarities and distinctions. A significant benefit of our dataset is its comprehensive transcription, rigorously annotated for stutters based on defined guidelines. This is distinct from other datasets that primarily focus on binary stutter event detection for uniform-length audio segments.

In summary, the notable diversity across participants and across speech types enriches our dataset, capturing a broad spectrum of stuttering patterns. Such diversity is advantageous as it provides a more comprehensive understanding of Chinese stuttering behaviors. Our analysis on the initial dataset provided us confidence that the AI models trained on this dataset can capture a broader range of speech patterns, enhancing their performance during real-world interactions. Additionally, when used for evaluation, our dataset offers a rigorous test bed, reflecting diverse scenarios and minimizing potential biases in model assessment.

7 DISCUSSION

7.1 Benefits of community-led data collection

The prevailing AI data paradigm frequently neglects marginalized communities, leading to leading to datasets that fail to accurately capture their nuances, and therefore AI applications don't cater to their specific needs [3, 8]. Compounding this issue are often exploitative data collection methods that raise psychological, trust, and privacy issues, further alienating and disempowering these communities [26].In the face of such systemic challenges, our case study on StammerTalk, a grassroots community in China for individuals who stutter, offers a promising alternative. This research highlights the advantages of community-led data collection and curation threefold. First, it showcases that both data collectors and contributors benefit from a profoundly positive experience, acquiring invaluable skills, insights, and emotions from the process. Second, this approach strengthens the bond and fosters a greater sense of belonging within the community, empowering its members. Lastly, this community-driven approach produces data of exceptional quality, characterized by diverse, authentic, and representative data. With the added richness from diverse participants and their wide-ranging speech patterns, the dataset offers a comprehensive representation of diverse stuttering behaviors.

7.2 The urgency to develop adequate socio-technical infrastructure for community data stwardship

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

7.2.1 **Open-sourcing datasets.** Open-sourcing datasets is inherently more complex and fraught with challenges 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 personal speech patterns.

The intrinsic value of our dataset lies in the unique speech characteristics of the individuals included. These distinct patterns are what make the dataset invaluable for research and AI model training. 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.

7.2.2 **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), and data processors (usually service providers or other companies). 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.

However, these models fail when the lines blur — when data subjects and controllers are essentially the same, or when the data controller isn't a traditional legal entity like a company. For instance, StammerTalk, being an unregistered grassroots community, doesn't fit neatly into any of these categories. The dilemma of crafting a participant agreement form for such a community is unavoidable. Ultimately, a temporary solution was adopted where a few StammerTalk members were designated as the primary legal parties. This is, however, far from an ideal representation of the community's intentions and creates disproportionate legal liabilities for the designated members. Given the fluid nature of virtual communities, there's significant turnover, with members periodically becoming inactive or leaving entirely. Such an arrangement could become problematic in the long run, highlighting the pressing need for more flexible and inclusive data governance frameworks.

7.2.3 **Geopolitical Complexities in Cross-sector Collaboration on AI Data.** The collaboration between StammerTalk and its diverse set of academic, industry, and nonprofit partners across China and the US added an extra layer of intricacy to the project, especially when working with personal data in a politically charged climate.

The current tension between the US and China in technological innovations, especially AI technologies, has triggered concerns and risks for partner organizations to engage and support this project. The ambiguity of defining "AI" from US government's perspective posed challenges. Queries such as "does it involve neural networks?" were raised. While there's room for exceptions in the realm of "basic AI research," determining what constitutes "basic" remains a grey area. The deliberate vagueness of these terms remains a question. A substantial amount of resources and legal expertise were required to establish clearance and legal protections for collaborating organizations to enter the partnership.

Concurrently, policy and legal challenges arise from the Chinese side, primarily due to the evolving landscape of data privacy laws. China recently introduced stringent personal data privacy laws targeting "companies" (including nonprofit organizations). The vagueness of these regulations and their potential application has sparked debates and uncertainties. While bigger companies may have the resources to comprehend, negotiate and navigate the laws, grassroot communities, such as StammerTalk, were left in a vulnerable position, unable to shoulder potential legal and financial risks.

7.2.4 **Navigating Cross-Border, Multinational Personal Data Collection and Protection Laws with Grassroots Online Communities.** Besides the geopolitical complexities, 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 diverse geographical spread of the community means traversing 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.

7.3 Conditions for successful community-led data collection

A successful community-led data curation project, like the one spearheaded 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?

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, we aim to inspire and provide a blueprint for other communities eager to initiate similar ventures.

8 LIMITATIONS AND FUTURE WORK

Our work comes with several limitations that require future investigations.

- **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.
- **Geographical Representation:** StammerTalk predominantly represents the stuttering community based in China. Although our data collectors, Rong and Lezhi, are based in the EU and U.S., the majority of our data contributors reflect Eastern Asian cultures and perspectives. Venturing further, it would be intriguing to see how this initiative evolves in regions with a higher level of awareness and acceptance towards stuttering, such as the U.S. and EU. Assessing community dynamics and data contribution processes in these contexts would add a new dimension to our understanding.
- **Utilizing the Dataset:** An evident next step for us is to delve deeper into the dataset we have acquired. We aim to undertake benchmarking experiments and harness the potential of this dataset to fine-tune state-of-the-art models, enhancing their performance and inclusivity.

9 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 exploration of community-led data collection using StammerTalk, an online community for Chinese-speaking people who stutter, as a case study. We found that grassroots initiatives can produce authentic and high-quality datasets, driven by intrinsic motivations that foster meaningful contributions and community connections. Participants derived empowerment, personal skills, and camaraderie from the process, illustrating the huge 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 global geo-political tensions and data regulations. It's imperative for stakeholders – ranging from industries to academia and policymakers – to recognize and invest in building robust infrastructures that empower marginalized communities in shaping their data narratives and AI-driven 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] 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.

[3] 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.

[4] 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

[5] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. " Why would anybody do this?" Understanding Older Adults' Motivations and Challenges in Crowd Work. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2246–2257.

[6] 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

[7] 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.

[8] Leigh Clark, Benjamin R Cowan, Abi Roper, Stephen Lindsay, and Owen Sheers. 2020. Speech diversity and speech interfaces: Considering an inclusive future through stammering. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.

[9] Marianne Dee and Vicki L Hanson. 2014. A large user pool for accessibility research with representative users. In *Proceedings of the 16th international acm sigaccess conference on computers & accessibility*. 35–42.

[10] Xianghua Ding, Patrick C Shih, and Ning Gu. 2017. Socially embedded work: A study of wheelchair users performing online crowd work in china. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 642–654.

[11] 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.

[12] Peter Howell, Stephen Davis, and Jon Bartrip. 2009. The university college london archive of stuttered speech (uclass). (2009).

[13] 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.

[14] 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*.

[15] 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.

[16] Rie Kamikubo, Utkarsh Dwivedi, and Hernisa Kacorri. 2021. Sharing practices for datasets related to accessibility and aging. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.

[17] Rie Kamikubo, Kyungjun Lee, and Hernisa Kacorri. 2023. Contributing to Accessibility Datasets: Reflections on Sharing Study Data by Blind People. 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 827, 18 pages. https://doi.org/10.1145/3544548.3581337

[18] 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

[19] 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.

[20] 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

[21] 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

[22] Qisheng Li, Krzysztof Z. Gajos, and Katharina Reinecke. 2018. Volunteer-Based Online Studies With Older Adults and People with Disabilities. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 229–241. https://doi.org/10.1145/3234695.3236360

[23] Qisheng Li, Josephine Lee, Christina Zhang, and Katharina Reinecke. 2021. How Online Tests Contribute to the Support System for People With Cognitive and Mental Disabilities *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 15 pages. https://doi.org/10.1145/3441852.3471229

[24] 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.

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[26] 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

[27] 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.

[28] 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.

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

[30] 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

[31] 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

[32] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63.

[33] Ana Rita S Valente, Luis MT Jesus, Andreia Hall, and Margaret Leahy. 2015. Event-and interval-based measurement of stuttering: a review. *International Journal of Language & Communication Disorders* 50, 1 (2015), 14–30.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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).

[39] 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