

“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/XXXXXXXX.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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Association for Computing Machinery.

Manuscript submitted to ACM

53 disabilities as *subjects* and *contributors* rather than the *owner* and *steward* of the collected data. The expert-led data
54 model deprives people with disabilities with their agency and control in the data collection process about them, leads to
55 challenges to engage and incentivize the community to participate [26].
56

57 The emerging practice of grassroots data collection presents opportunities for marginalized communities to take
58 initiative and control over their data experiences [1]. In this paper, we present a case study on grassroots data collection
59 initiative led by StammerTalk, an online community for Chinese-speaking people who stutter (PWS). This initiative
60 aimed to create and curate one of the first Chinese stuttered speech datasets to improve their experience with speech
61 AI technologies. By closely following the process from its inception by the StammerTalk community, we collected
62 ethnographic data about this process through observations, interviews, and survey with community members, to
63 understand the process, benefits, and challenges for community data stewardship. Our research questions are:
64

- 65 • RQ1: What are the motivation, experience, and challenges of the community members involved in community-
66 led data collection and how they differ from expert-led practice?
67
- 68 • RQ2: What are the characteristics of the result dataset and how it differs or resembles similar dataset collected
69 by experts?
70

71 Our study shows that the community-led data stewardship model not only produce high-quality dataset that
72 authentically represent the community in AI data, but also benefits the community and its members beyond the tangible
73 technical outputs. Contrary to the findings from previous research [26], community members who participated in the
74 community-led data collection were driven by intrinsic goals - such as the making meaningful contribution to the
75 community and connecting with other community members in collective actions, rather than monetary compensation.
76 Consequently, community participants had positive experiences during the data collection, appreciating the empathy,
77 understanding, and personal connections provided by community members who administered the data collection.
78 Beyond the enjoyable experiences, community participants also gained valuable skills, broadened perspectives on
79 stuttering, a sense of empowerment, and stronger bonds that could lead to long-term growth for the community. Our
80 technical evaluation of a subset of the collected data suggests the efficacy of the community-led data collection to
81 produce more diverse and authentic dataset than the traditional approach.
82

83 On the other hand, our study also uncovers challenges the community faced, due to limited resources and current
84 socio-technical infrastructure for personal data. Besides the time and energy required for community members to
85 design, execute, quality control the data collection process, they also face structural challenges introduced by existing
86 socio-technical infrastructures that were inadequate for community data stewardship. As a result, the community
87 need to navigate complex geo-political tensions and cross-border data regulations when working with its partners and
88 members across the globe.
89

90 Taken together, our work illustrates the benefits and challenges of community data stewardship for AI data, and calls
91 for the attention and investment from industry practitioners, academic researchers, and policy makers to develop the
92 socio-technical infrastructure needed in order for marginalized community to take charge of their data and data-driven
93 experiences.
94

95 2 RELATED WORK

96 2.1 Bias and Under-representation in AI Model Training

97 A growing body of research showed the under-representation or misrepresentation of marginalized communities
98 in AI training datasets. For instance, Buolamwini and Gebru highlighted the under-representation of women and
99

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

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

125 The research emphasized the urgency of considering AI's impact in broader contexts. It is crucial to offer users avenues
126 to rectify errors and ensure the inclusion of marginalized groups, including people with disabilities, in data sourcing
127 and model testing phases. This inclusion is vital for developing truly inclusive and representative AI applications.
128

129 **2.2 Data Practice in HCI and Accessibility**

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

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

143 Beyond paid platforms, some research efforts rely on volunteer-based data collection platforms, like LabintheWild [22,
144 23]. Rather than financial compensation, participants are incentivized through personalized feedback. Such feedback
145 serves multiple purposes: from allowing self-comparison with others to facilitating word-of-mouth recruitment as users
146 share their results on social networks. Importantly, online tests have been found to reduce barriers to professional
147 diagnosis for participants, and can help shape their disability identity, fostering self-acceptance. However, a gap remains
148 in terms of two-way communication between participants and researchers, with participants expressing the desire for
149 more interactive feedback and support [22].

152 Another collective body of work underscores the essential role of accessibility datasets in driving AI applications
153 tailored to the needs of disabled communities. Recognizing the pivotal role of obtaining data from disabled communities,
154 Park et al. delved into designing an online infrastructure tailored to collect AI data from this demographic. Their
155

157 study outlined essential guidelines, emphasizing motivation strategies, transparent communication, and accessibility
158 considerations to ensure inclusive AI datasets [26]. Additionally, Kamikubo et al. (2021) offers a broad view, examining
159 the balance of risks and benefits in sharing accessibility data over three decades, highlighting the scarcity of datasets
160 directly sourced from underrepresented communities [16]. Complementing this, both Theodorou et al. and Kamikubo
161 et al. (2023) delve deeper into specific community interactions, the former focusing on blind or low vision participants'
162 contributions for teachable object recognition [31], and the latter exploring the decision-making processes of blind
163 individuals regarding data sharing, factoring in various modalities, objects, environments, demographics and access
164 conditions [17]. Furthermore, Danielle et al. adds a cultural layer to this discourse, addressing the sensitivity of sign
165 language datasets rooted in the unique identity of the Deaf community and the historical implications associated with
166 it [4].
167

168 While these studies involve the community in data collection and considerations, none focused on the grassroots
169 aspect our research does – where the community is not just a participant but is intrinsically involved as the steward,
170 initiator, and ultimate decision-maker.
171

172 2.3 Grassroots Data Stewardship Models

173 While considerable research has explored various aspects of data practices, little attention has been devoted to grassroots
174 data initiatives. grassroots data practices are especially potent in settings where mainstream avenues fail to cater to
175 local necessities. A case in point is the local data practices of Mexico City citizens who harness Facebook for crime
176 reporting and safety information dissemination. Garcia et al. studied these communities and found that administrators
177 play a pivotal role, using crowdsourced data to engage with various stakeholders [1]. Their methods of data gathering,
178 curation, and dissemination amplify trust and foster extended participation, establishing grassroots data as a key tool
179 for tackling unreported safety concerns. This work underscores the role of social media in enabling political action
180 where traditional institutions falter.
181

182 However, contemporary data practices and stewardship models are insufficient to provide a comprehensive represen-
183 tation of marginalized communities in AI data. Janis Wong's overview suggests that company-driven data collection
184 and stewardship are riddled with challenges, stemming from a lack of trust, inadequate incentives, and a lack of agency
185 within communities [35]. In the quest for alternative models, the indigenous data sovereignty initiatives undertaken in
186 Australia and Hawaii offer promising insights [34]. Drawing inspiration from such endeavors, our work aspires to build
187 a community-led data model, specifically catering to the stuttering community.
188

189 While the participatory data stewardship model stands as a benchmark in the realm of data collection and manage-
190 ment [29]¹, our methodology endeavors to transcend its boundaries. In our paradigm, the community transcends the
191 traditional role of mere participants; they emerge as the initiators and leaders, emphasizing a profound shift in data
192 stewardship dynamics.
193

194 2.4 Stuttering and Speech AI

195 Individuals who stutter often confront evident performance discrepancies when interacting with current ASR-enabled
196 applications. While recent technical advancements have shown potential in enhancing ASR models for stuttered speech,
197 the scarcity of stuttered speech data remains a significant barrier to progress.
198

199
200
201
202
203
204
205
206
207 ¹<https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/>

209 **2.4.1 ASR's Inadequacy for Individuals Who Stutter.** Bleakley et al.'s recent user experience research illuminates
210 the pronounced challenges faced by people who stutter when interacting with smart speakers [3]. Not only do
211 these individuals encounter inaccurate transcriptions, but ASR systems often interrupt them mid-speech or, at times,
212 completely fail to respond to their requests. Such behavior is not merely a technical glitch; it effectively imposes
213 structural barriers that restrict access to essential services and information for those who stutter. Echoing this sentiment,
214 Clark et al. emphasize the necessity of accommodating diverse speech patterns in ASR development and outline the
215 fundamental challenges tied to creating truly inclusive speech interfaces [8].
216

217 Highlighting the urgency of these findings, Apple's research team, led by Lea *et al.*, recently conducted benchmarks
218 on their ASR systems specifically for stuttered speech. Their results unveiled a stark performance discrepancy, with a
219 tenfold increase in Word Error Rate (WER) for users with severe stuttering. Moreover, when evaluating the performance
220 of endpointer models in terms of truncation rate, the disparity remains evident: a significant proportion of utterances
221 from people who stutter are prematurely cut off compared to those from the general population [20]. This finding is
222 consistent with previous observations by Lea *et al.*, further emphasizing the pressing need for advancements in this
223 domain [21].
224
225
226
227
228
229

230
231 **2.4.2 Existing technical efforts.** The endeavor to enhance ASRs for stuttered speech has been primarily characterized
232 by two main thrusts: i) the establishment and curation of stuttered speech datasets, and ii) the modification of pre-existing
233 ASR models to improve their proficiency in detecting and understanding stuttered speech.
234

235 **i) Datasets of Stuttered Speech.** With some originally conceived as tools for speech therapy, several datasets
236 featuring speech samples from individuals with stutters have been compiled. These include the FluencyBank [28],
237 the University College London's Archive of Stuttered Speech (UCLASS)[12], the SEP-28K dataset sourced from public
238 podcasts[21], non-standard speech from Project Euphonia [24], and the LibriStutter dataset that comprises synthesized
239 stutters [19]. Despite the substantial size of these datasets, they contain only a limited portion of stuttered speech
240 samples. Furthermore, inconsistencies in stutter-specific annotations present challenges, both of which hinder the
241 effective training of advanced ASR models for stuttered speech.
242

243 **ii) Adapting Existing ASR Models for Atypical Speech.** Google has been at the forefront of these efforts via
244 Project Euphonia [24], making significant strides in refining their existing model using a relatively modest volume of
245 atypical speech data. However, their research has predominantly concentrated on recognizing brief phrases (such as
246 command words) and non-stuttered speech forms, such as speech patterns of people with ALS or Down Syndrome.
247 As a result, the transferability of their advances to contexts involving conversational and/or stuttered speech remains
248 uncertain. In addition to releasing the Sep-28K dataset [21], Apple has also explored the tuning of ASR models for
249 dysfluent speech. Lea *et al.* have recently compared three interventions to augment ASR systems, some of which require
250 minimal stuttered speech samples [20].
251

252 In summary, despite the increasing interest from major technology companies in addressing this challenge, progress
253 has been gradual, and the existing gap remains. The challenges faced by individuals who stutter are manifold, encom-
254 passing both personal and systemic dimensions. It's imperative to take actions to ensure AI systems don't further
255 exacerbate disparities or suppress certain demographics.
256
257
258
259
260

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 WeChat² 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.

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

313 3.2.1 **Conception.** In December 2022, the concept of gathering a Chinese language stuttered speech dataset emerged
314 during a discussion in a WeChat group chat between the StammerTalk leadership and one author of this paper. The
315 primary motivation was to utilize this dataset as benchmarking resources to enhance ASR models' performance for
316 stuttered speech. While the proposal was met with enthusiasm and appreciation from all group members, Rong, a key
317 member of the StammerTalk leadership, graciously stepped forward to spearhead the project's planning and execution.
318
319

320 3.2.2 **Preparation.** To set the groundwork for data collection, the StammerTalk core team undertook several pivotal
321 steps:

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

336
337
338
339
340 3.2.3 **Participant recruitment.** Meanwhile, StammerTalk began its participant recruitment for the data collection
341 through a detailed post on its WeChat official account. The main contents of this recruitment post included:

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

354
355
356
357
358 The initial recruitment effort attracted over 40 community members. A subsequent recruitment phase, using the
359 same strategy, was initiated in July 2023.

360
361 3.2.4 **Data collection.** Upon signing up for a data collection session, community members were presented with a
362 participant agreement form. This form outlined the purpose of the data collection, potential applications of the collected
363
364

365 data, privacy protection measures, and opportunities for participants to be involved in data management. Once the
366 form was acknowledged and signed, participants were scheduled for a structured 90-minute data collection session
367 conducted via Zoom or Tencent Meet, broken down as follows:
368

- 369 (1) **Preparation (30 mins)**: The session commenced with an introduction from the StammerTalk core team member
370 (the interviewer) hosting the meeting. The interviewer then briefed the participant (the interviewee) on the
371 session's structure. Additionally, the interviewer ensured optimal technical and environmental conditions for
372 the recording. Participants were then requested to complete the Overall Assessment of the Speaker's Experience
373 of Stuttering (OASES) [37], the results of which were paired with the recordings as metadata.
- 375 (2) **Unscripted Spontaneous Conversation (30 mins)**: A casual conversation took place, focusing on the
376 participant's background and personal stuttering experiences.
- 377 (3) **Voice Command Recitation (30 mins)**: Participants were provided a set of typical voice commands to read
378 aloud.
379

380 The latter two components of the session were audio recorded. Subsequently, these recordings were securely stored
381 in a shared Google Drive folder, accessible only to the StammerTalk core team and selected partners. Approximately an
382 hour of speech data was collected from each session. Trained annotators executed the speech-to-text transcription and
383 stuttering annotations in 10-hour batch intervals.
384

385 By August 2023, a total of 58 StammerTalk community members had participated in this data collection process. For
386 the rest of this paper, we will refer to the StammerTalk core team members who collected and processed the data as **data**
387 **collectors**; the community members who participated in the recording sessions are denoted as **data contributors**.
388
389

390 3.3 Positionality Statement 391

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

394 Both of us are Mandarin-speaking, Asian/Asian American women residing in North America. Together, we bring 22
395 years of experience working in academia and the corporate, with expertise in data science, HCI, accessibility, and AI.
396 While affiliated with technology companies and/or university research institutes, we both had experience gathering
397 data from individuals with disabilities, both directly through company's or lab's platforms, and indirectly via data
398 vendors.
399

400 One of us identifies as a person who stutters. This author has engaged with StammerTalk, attending a self-help
401 session and being interviewed for their podcast. Additionally, she has personal ties with the StammerTalk moderators
402 and collaborated on other stuttering-related projects beyond this data collection endeavor.
403

404 Though our close relationship with StammerTalk and shared experiences as stutterers fostered trust and commu-
405 nity access, it didn't entirely negate the power dynamics between researchers and subjects. Our socioeconomic and
406 educational backgrounds also granted us certain privileges relative to many community members we engaged with.
407
408

409 4 METHOD 410

411 4.1 Semi-structure interview with data collectors

412 We conducted semi-structured interviews with the two primary data collectors of this initiative. Our goal was to delve
413 deeper into their motivations, capture their experiences, and understand the challenges and insights they garnered as
414 leaders throughout the data collection journey.
415
416

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

469 and to ensure we received a holistic view, we designed a survey for these individuals. This methodological expansion
470 aimed to validate and potentially enrich the themes that emerged from the moderator interviews, ensuring their
471 relevance and resonance with the larger data contributors pool. By consolidate perspectives from both data collectors
472 and data contributors, we sought to capture a comprehensive landscape of the data collection process and its nuances.
473 The survey was conducted in Mandarin, and the results presented in subsequent sections are translated into English.
474
475

476 **Survey Questions.** The survey design was informed and shaped by insights garnered from prior interviews with data
477 collectors. The survey comprised 14 distinct questions, both open- and closed-ended. For a comprehensive view of the
478 entire survey, please refer to the Supplementary Material. The survey can be categorized into the following segments:
479

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

495
496
497
498
499 Through this structured approach, the survey was designed to comprehensively capture data contributors' experiences,
500 challenges, and insights.
501

502
503 **Recruitment.** Data contributors were individually invited by Rong, one of the data collectors, to complete the survey.
504 They were informed that the survey was administrated by [Organization Name], designed to better understand and
505 improve the data collection process, and they were be compensated with ¥30 RMB (approximately \$5 USD) upon
506 completion of the survey. The survey was hosted through Tecent Survey platform. The survey took about 5 minutes
507 per respondent, and compensation were distributed by Rong on behalf of [Organization Name] to the respondents
508 through WeChat Pay.
509
510

511 **Analysis.** Among all the 58 data contributors who completed the data collection sessions by the time we administrated
512 the survey, 55 people (95%) submitted their responses to the survey. The mean survey completion time was 5 minutes.
513

514 For open-form questions, we utilized an iterative coding methodology [13]. For each question, one author developed
515 an initial codebook. Two authors then collaboratively discussed and refined the codebook, applying it iteratively to all
516 responses.
517

518 To analyze the qualitative data, we focused on descriptive statistics, primarily frequencies. Given the nature of our
519 survey, which aimed to understand holistic experiences rather than identifying correlations between variables, most
520

521 questions were of the 'apply-all-that-apply' type. Thus, complex statistical analyses were not deemed appropriate or
522 necessary for our research objectives.
523

524 **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%)
525 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,
526 while the other 20 people (36.4%) identified as female. Our data contributors have a wide range of occupations: a
527 significant number of participants (23.6%) identified as students; other notable occupations include IT-related roles
528 (11%), medical professionals (7%), public service roles (e.g., civil servants, teachers), and roles in various specialized
529 fields ranging from energy sectors to biotechnology.
530
531

532 The majority of our participants (83.6%) also indicated that they have received some form of stutter-related support in
533 the past, with the types of support not being mutually exclusive. Specifically, 25 participants had undergone stuttering
534 therapy or training, 27 had attended online or offline stuttering self-help groups, another 27 identified as members of
535 online or offline communities for people who stutter, such as the StammerTalk WeChat group or National Stuttering
536 Association (NSA) in the U.S., and 17 had participated in stuttering-related community events like lectures or public
537 activities. Conversely, 9 individuals (16.4%) reported not having engaged in any of the aforementioned forms of support.
538
539

540 5 FINDINGS

541
542 Here we describe the major findings from our work, centering around the incentives, experiences, gains, and challenges
543 for community members to lead and participate in the data collection process. Our findings highlight that, contrary
544 to what is reported in previous research [26], StammerTalk members who participated in the community-led data
545 collection were driven by intrinsic incentives - such as the making meaningful contribution to the community and
546 connecting with other community members, rather than monetary compensation. Community members also gained
547 empathy, understanding, knowledge, and personal connection with each other during the data collection, resulting in
548 overwhelmingly positive experiences and a sense of self and community empowerment.
549

550 Our data also uncover the challenges for community-led data collection, namely, the significant time commitments,
551 the resources required to annotate the recorded speech data, and the uncertainties with legal and privacy implications.
552 While the StammerTalk community was pragmatic and resourceful to navigate these challenges, our study calls for
553 the development of adequate socio-technical infrastructure for a broader and easier adoption of community data
554 stewardship model from other marginalized communities.
555
556

557 5.1 Incentives

558
559 The StammerTalk community's primary drive for the stuttered speech collection project stemmed from **intrinsic**
560 **motivations** such as community empowerment and forging interpersonal connections, overshadowing external
561 incentives like monetary rewards.
562

563 Both data collectors, Rong and Lezhi, have backgrounds in technology and aimed to use their skills to address the
564 community's technological challenges. Rong, associated with a commercial speech technology company and leading
565 this project, emphasized:
566
567

568 *I'm professionally involved in this space, understanding the entire process well... The various stages like*
569 *data collection, annotation, model training, and evaluation are part of my daily jobs. Therefore, undertaking*
570 *this project end-to-end would be intriguing and meaning for me. (Rong)*
571

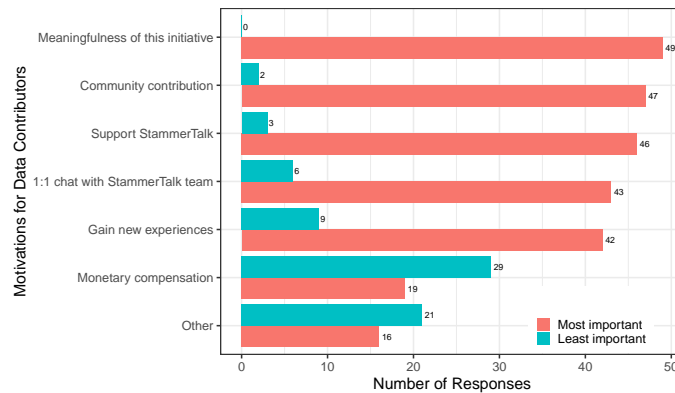


Fig. 1. The most and least important reasons for data contributors to participate in data collection project.

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)

625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676

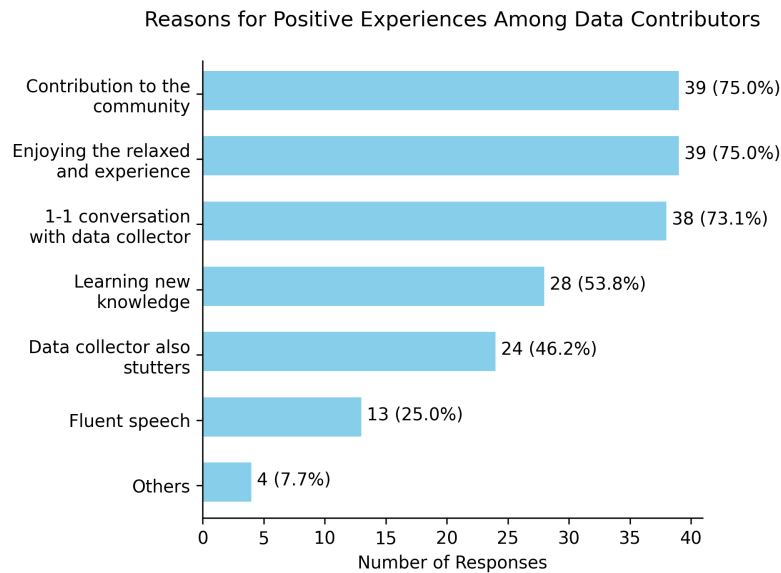


Fig. 2. The primary reasons that led to the positive experiences among the data contributors in the data collection project.

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

677 contributing to their feelings. The summarized responses can be found in Fig. 2. The three leading reasons contributing
678 to the positive experiences of data contributors were: a sense of making a meaningful contribution to the community
679 (75%, N=39/52), a relaxed and comfortable atmosphere during the interview (75%, N=39/52), and the unique experience
680 of having a one-on-one conversation with another person who also stutters (73%, N=38/52). These results resonate
681 with our earlier findings regarding the primary motivations for participation, highlighting the value of community
682 engagement and fostering deeper connections among its members for the data contributors.
683
684

685 **5.2.2 Positive Experiences of Data Contributors - Role of Empathetic Data Collectors.** Most data contributors
686 found their interaction with data collectors during the data collection process uniquely positive, greatly contrasting
687 with their typical speaking experiences. The ambiance of trust and comfort during the interviews largely stemmed
688 from the empathetic nature of the data collectors.
689

690 A significant number of data contributors (N=23/54) appreciated being interviewed by someone who also stutters.
691 As Rong observed, the mutual experience of stuttering established an immediate sense of trust. The participants
692 often remarked, “ *oh, you also stutter!*”, followed by, “*now I can relax.*” Lezhi had similar observations, “ *People who
693 stutter usually engage in a psychological defense when it comes to speaking,(...) Since my stuttering is relatively severe, the
694 participants might feel there is nothing they need to hide when speaking with me.*”
695
696

697 Beyond this shared identity, the data collectors took deliberate steps to nurture a positive environment. They shared
698 personal experiences, adapted conversation topics to accommodate interviewee’s mental and emotional state, and
699 avoided any behavior that might be perceived as intimidating. Lezhi remarked on the importance of this approach,
700 emphasizing her ability to discern when a participant might be feeling nervous and adjusting her approach accordingly:
701

702 *(...) When someone was nervous, I would chose to ask them some easy topics to help them relax. (...) As a
703 person who stutters, I know what types of topics will make them more nervous, I could also quickly identify
704 the characteristics of their stutter and which words might be difficult for them to say. (Lezhi)*
705
706

707 Survey results reaffirmed that the data collectors have successfully connected with, emphasized, and accommodate
708 the data contributors during and after the interview. Of 55 respondents, 54 rated their interaction with the data collectors
709 as either “*Good*” or “*Very Good*.” As shown in Fig.3, respondents particularly valued the data collectors’ attentive listening
710 (N=51/54), clear communication about the data collection process (N=45/54), and the substantial empathy shown by the
711 interviewers (N=38/54).
712

713 Interestingly, the supportive environment meant some data contributors spoke more fluently than usual during the
714 recording session, with N=13 out of 54 citing this increased fluency as a factor in their positive experience. While all
715 participants identified as people who stutter, some were still self-conscious about it and naturally aimed for fluency. Such
716 a focus on fluency might be attributed to the physiological and emotional challenges associated with stuttering [36].
717 However, this increased fluency could result in the collected data being less representative of typical stuttering patterns.
718 This potential limitation will be explored further in the “Challenges” section below.
719
720

721 722 723 **5.3 Gains**

724 Beyond the direct benefit of curating an immediate, tangible data asset for the community, participants of this project
725 also gained valuable skills and experiences, broadened perspectives, and connections that could lead to long-term
726 positive outcomes for the community.
727

729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780

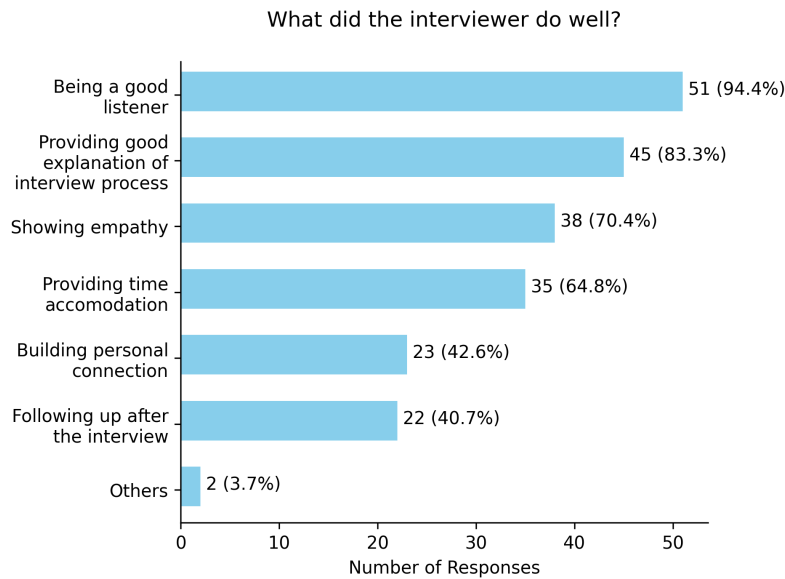


Fig. 3. Data contributors' feedback on data collectors' competencies during the data collection project.

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."

781 Some participants (N=21) shared that participation in the data collection project strengthened their feelings of unity,
782 recognition, and empowerment within the stuttering community, fostering a deeper sense of belonging and collective
783 progress. One participant expressed (P19), “ [I love] meeting more friends and teachers. It made me realize that there
784 are many people in the world just like me. We all strive to live well, working hard to overcome the impact of stuttering on
785 ourselves.” Others (e.g. P20), acknowledged the broader awareness and understanding brought about by the project: “ I
786 realized that there are so many people continuously paying attention to the stuttering community... leading more people who
787 stutter to focus on themselves.” This growing unity and recognition, as summarized by another participant P1, has led to
788 a feeling that “ our community has united and received more attention, advancing the progress of stuttering treatment
789 in China.” Collectively, these reflections emphasize the transformative potential of community-driven initiatives in
790 fostering a deeper sense of belonging and empowerment within the stuttering community as well as within broader
791 societal frameworks.
792

793 Participants (N=14) also gained immensely from the genuine, one-on-one communication opportunities provided to
794 them. Their feedback consistently underscores the immense personal growth, realization of their inherent potential,
795 and emotional relief they experienced by having a space where they could “ communicate with others and open one’s
796 heart ” and “ have the courage to face the real self and accept one’s stuttering behavior.” Free from judgment and without
797 the burden of hiding their stutter, they felt a profound sense of liberation and empowerment. Engaging with someone
798 from “ a similar group ” deepened this transformative experience, accentuating the power of shared experiences and
799 the realization of one’s true potential. As P33 expressed, being able to “ freely express without consciously hiding my
800 stutter ” not only served as a medium of self-expression but also as an affirmation of self-acceptance and self-worth.
801 The understanding and respect they gained, especially from an “ interviewer who also stutters,” instilled a sense of hope
802 and fueled their optimism. Through these interactions, it becomes clear that participants received not just an avenue to
803 express, but also deeper self-awareness, acceptance, and a renewed sense of empowerment — driven by the profound
804 empathy and connection they experienced.
805

806 Other participants (N=10) say that the biggest gain from participating the data collection project is having learned
807 new knowledge about stutter. For instance, P9 mentioned “ I learned that one can approach stuttering from a scientific
808 perspective.” Others emphasized the learning gained uniquely from talking to people who also stutter. As P45 put it: “
809 The interviewer’s pronunciation and manner of speaking in a very slow and gentle voice slightly improved their speech
810 fluency – it effectively conveyed their message to me. This deeply resonated with me, and I am currently learning this way
811 of speaking.”; similarly, P8 mentioned: “ The biggest takeaway was being interviewed by someone who stutters like me
812 and conversing with them. They briefly shared some methods on how to alleviate awkwardness and how they cope with
813 stuttering, as well as how to relieve anxiety. Through our conversation, I learned quite a lot.”
814

815 In summary, the data contributors greatly valued their participation in the data collection project for the interpersonal
816 connections, empowerment, and advocacy it fostered. Similarly, data collectors experienced personal growth and formed
817 lasting connections, emphasizing the project’s profound impact beyond its primary objective.
818

819 5.4 Challenges

820 Despite the community members’ strong motivation and positive experiences, some substantial challenges are un-
821 avoidable during the process. While the StammerTalk community had managed to come up with creative strategies to
822 navigate these challenges, some questions remained open as the project moves forward.
823

824 **5.4.1 Challenges for data collectors.** There were four major challenges data collectors had to navigate through.
825

833 **(1) Time Commitment.** First, the time commitment has been one of the biggest challenges to make progress at
834 a desired pace. Rong and Lezhi, both employed full-time, could only allocate evenings and weekends to the project,
835 affecting their personal and family time. Given their locations in Europe and the US, coupled with the primary participant
836 base in China, scheduling was particularly challenging due to time zone differences. Their limited availability and
837 the often last-minute rescheduling requests resulted in, at most, one or two recording sessions per week. Given the
838 significant amount of time and energy required to recruit and schedule for the data contributors, and the fact that “*for*
839 *the second and the third batch, there would be probably fewer people signing up,*” Rong anticipated it might push the
840 project’s completion timeline to potentially over a year to reach the goal of 100 hours from 100 individuals.
841

842 **(2) Data annotation.** Second, as briefly introduced in the *Background* section, finding annotation services to
843 accurately annotate the collected Chinese stuttered speech sample was also challenging, as it had never been done
844 before at this scale. As a result, Rong had to spend substantial amount of time and energy to create detailed annotation
845 guidelines and to train the annotators, who were non-stuttering and had no prior experience of annotating stuttered
846 speech. It took three iterations for the annotators to be able to identify and label the stuttering events. During each
847 iteration, Rong would carefully review the annotations produced by the annotators, and returned with corrections with
848 detailed explanations. While the entire process was tedious and time consuming, Rong recognized the dedication of the
849 annotators and their adaptability, but also realized that, due to the pro bono nature of the service, achieving the ideal
850 annotations consistent with stuttering professionals was ambitious:
851

852 *It took the annotators quite a lot of efforts during our training. Since none of them stutters, nor did they*
853 *work with PWS professionally, it is very difficult for them to produce the consistent annotations as stuttering*
854 *professionals do. After three iterations, although there were still some places that were unsatisfactory to*
855 *me, I thought it was already very good for non-stuttering annotators to have this level of quality in their*
856 *annotations.* (Rong)
857

858 **(3) Data quality and representativeness.**

859 Another key challenge faced by the data collectors was ensuring both the quality and representativeness of the
860 recorded speech. They aimed to balance between capturing clear sound, diverse speech types, and varying stuttering
861 patterns, sometimes at the cost of the positive experience of the data contributors.
862

863 Concerning **sound quality**, although data contributors received guidelines on environmental and technical settings,
864 not all complied. For instance, Lezhi encountered situations where contributors were in noisy surroundings or interrupted
865 by phone calls, necessitating either waits or rescheduling to achieve optimal sound conditions.
866

867 The data collectors also strived to have the data sufficiently cover **the variety in stuttering patterns and severity**
868 **levels**. Stuttering, being multifaceted, varies in frequency, severity, and manifestation across individuals and contexts [20,
869 36]. The recording sessions –combining unscripted conversations with recitation of common voice commands – aimed
870 to capture different speaking contexts. However, the comfort ambiance often led to participants stuttering less than
871 usual, particularly during voice command recitation, which could limit the data’s real-world representativeness.
872

873 To address this issue, the data collectors employed strategies, such as 1) encouraging voluntary stuttering – imitating
874 stuttering on words they typically wouldn’t, and 2) posing challenging questions to induce tension.
875

876 While these strategies help increase the frequency of stuttering, there are trade-offs, such as the trade-off of tension
877 and openness during the unscript conversations. As Lezhi explained,
878

879 *There needs to be a balance. When someone was nervous, they could choose to speak less; when someone*
880 *was relaxed, they would not stutter. When someone was nervous, I would chose to ask them some easy topics*
881

885 *to help them relax; when someone was very relaxed, I would ask a less comfortable question. As a person*
 886 *who stutters, I know what types of topics will make them more nervous. (...) Based on what he (the data*
 887 *contributor) shared about his background, I would intentionally follow up with some additional questions*
 888 *make him feel like at a job interview, to create a bit more tension. (Lezhi)*
 889

890
 891
 892 Despite the lower-than-expected stuttering frequency, the data collectors believed their method best represented and
 893 empowered the stuttering community. Contributors weren't pre-screened for speech samples or stuttering descriptions.
 894 While they did complete the Overall Assessment of the Speaker's Experience of Stuttering (OASES) [37], it wasn't
 895 a selection criterion but rather treated as metadata. Rong reflected upon the selection criteria, and emphasized that
 896 a person's self-identification as someone who stutters should be the sole requirement for participation, making the
 897 sample intrinsically representative of stuttering community and eliminating potential biases. This approach accentuates
 898 the difference between community-led and third-party data collection. *Unlike commercial entities that might exclude*
 899 *someone for not being "disabled enough", community-led efforts, like this one, prioritize self-identity and inclusion.*

900 **(4) Data protection and governance.**

901 Ensuring data protection and governance posed a another notable challenge. Given that interviews delved deep
 902 into contributors' stuttering experiences, many participants shared intimate details about their lives. Such openness
 903 enhanced the data's authenticity but also raised privacy concerns. Though data collectors attempted to safeguard privacy
 904 by editing out personal details and encouraging pseudonyms, the effectiveness of these measures in fully anonymizing
 905 the dataset remains uncertain. While contributors consented to sharing their data for non-commercial uses, uncertainty
 906 around whether and how to share this dataset remains even among the community itself. Rong supports releasing the
 907 dataset under a non-commercial license, while Lezhi, citing legal and privacy concerns, believes only analyses and
 908 models derived from the data should be open-sourced. The complexities of global regulations surrounding biometric
 909 data, which includes speech, coupled with limited resources and expertise of StammerTalk being a grassroots online
 910 community, introduce potential legal liabilities for data collectors.
 911
 912
 913
 914
 915
 916
 917
 918

919 **Challenges for data contributors.** As shown in Figure 4, Of the 49 data contributors who responded to the survey
 920 question, 18 identified their stuttering during the interview as a challenge they faced. Another 13 participants felt
 921 nervous, an emotion that aligns with findings from the previous "Experiences" section where many contributors revealed
 922 they were still self-conscious about their stutter, and many desired fluency. This sentiment mirrors the often-felt physical
 923 tension and discomfort that many who stutter experience during speech. In a contrasting vein, 17 participants found it
 924 challenging to deliberately stutter during the interview, a finding that intriguingly mirrors the challenges faced by
 925 data collectors who sought to elicit a broader range of stuttering for data diversity. Some participants found the voice
 926 command reading tasks monotonous, especially since they had to repeat several commands multiple times. Additionally,
 927 11 contributors expressed a desire for more 1-on-1 interaction time with data collectors, underlining their interest in
 928 engaging and learning from community leaders and members.
 929
 930

931 In summary, we identified key obstacles in community-led data collection: substantial time commitments, the
 932 necessity of resources for speech data annotation, legal and privacy uncertainties, and finding the balance between
 933 stutter data representativeness and data contributors' comfortness.
 934
 935
 936

937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988

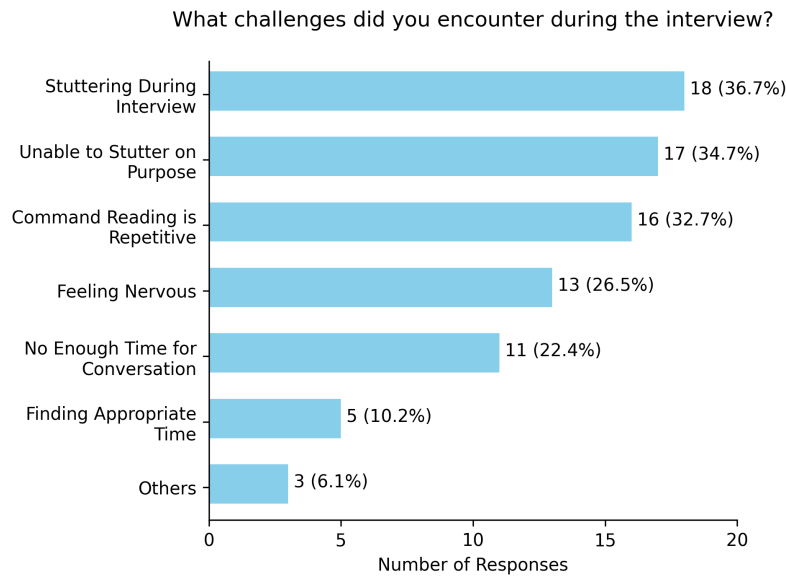


Fig. 4. Primary challenges faced by data collectors during the data collection process.

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 Praat³ 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 stewardship

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.

1249 Concurrently, policy and legal challenges arise from the Chinese side, primarily due to the evolving landscape of data
1250 privacy laws. China recently introduced stringent personal data privacy laws targeting “companies“ (including nonprofit
1251 organizations). The vagueness of these regulations and their potential application has sparked debates and uncertainties.
1252 While bigger companies may have the resources to comprehend, negotiate and navigate the laws, grassroots communities,
1253 such as StammerTalk, were left in a vulnerable position, unable to shoulder potential legal and financial risks.
1254
1255

1256 **7.2.4 Navigating Cross-Border, Multinational Personal Data Collection and Protection Laws with Grassroots**
1257 **Online Communities.** Besides the geopolitical complexities, the StammerTalk community also needed to navigate
1258 the multifaceted web of international data protection laws. Since the StammerTalk community solely exists online —
1259 holding meetings via Zoom and group chats, its members are distributed globally across geographical borders. The act
1260 of collecting data from community members thus becomes a cross-border undertaking. This results in the need to juggle
1261 multiple regulatory frameworks from regions such as the US, EU, and China, each with its nuances, and sometimes,
1262 contradictions.
1263
1264

1265 The diverse geographical spread of the community means traversing a labyrinth of legal guidelines, each with its
1266 unique stipulations. This complexity not only incurs considerable legal and procedural costs but also poses potential
1267 risks. Ensuring compliance with every relevant regulation becomes a daunting task, magnifying the exposure to
1268 potential legal liabilities that the community could not afford.
1269
1270

1271 **7.3 Conditions for successful community-led data collection**

1272 A successful community-led data curation project, like the one spearheaded by StammerTalk, is often influenced by
1273 a combination of factors. The question arises: which types of communities are best positioned to embark on similar
1274 initiatives? Alternatively, how can we better prepare communities to take on such initiatives?
1275
1276

1277 **Technical Expertise Within the Community:** A cornerstone of this project’s success was the technical proficiency
1278 present within the community. Rong’s professional background in speech AI technology endowed him with a thorough
1279 understanding of the complexity of the data collection process. His expertise not only influenced the initiative’s inception
1280 but also ensured that the necessary resources and steps were identified and followed.
1281

1282 **Resourcefulness:** An essential attribute for success is the ability to harness available resources effectively. This
1283 initiative was characterized by early partnerships and stakeholder buy-ins, ensuring access to pivotal assets such as
1284 annotation services.
1285

1286 **Reputation of Community Organizers:** The standing of the community organizers plays a pivotal role in the
1287 project’s overall reception and participation rates. When community members trust and respect the organizers, they
1288 are more inclined to participate. The positive reputation of the StammerTalk organizers created an environment where
1289 members were not only eager to engage but also looked forward to their interactions, keen on acquiring more knowledge
1290 and making meaningful contributions.
1291

1292 In summary, the success of such community-driven endeavors is multifaceted, requiring a blend of expertise, resources,
1293 and reputation. By maintaining transparency and openness throughout the project, we aim to inspire and provide a
1294 blueprint for other communities eager to initiate similar ventures.
1295
1296

1297 **8 LIMITATIONS AND FUTURE WORK**

1298 Our work comes with several limitations that require future investigations.
1299
1300

- 1301 • **Generalizability and Scope:** This study revolves around a specific case with a relatively small community
1302 leadership. We conducted interviews primarily with two members, which limits the breadth of our insights.
1303 While the findings provide valuable insights into StammerTalk, they may not be directly transferable to stuttering
1304 communities from other regions or other disability communities at large. Nevertheless, we hope our efforts
1305 serve as a catalyst, inspiring other communities to explore this domain with us.
- 1306 • **Geographical Representation:** StammerTalk predominantly represents the stuttering community based in
1307 China. Although our data collectors, Rong and Lezhi, are based in the EU and U.S., the majority of our data
1308 contributors reflect Eastern Asian cultures and perspectives. Venturing further, it would be intriguing to see
1309 how this initiative evolves in regions with a higher level of awareness and acceptance towards stuttering, such
1310 as the U.S. and EU. Assessing community dynamics and data contribution processes in these contexts would
1311 add a new dimension to our understanding.
- 1312 • **Utilizing the Dataset:** An evident next step for us is to delve deeper into the dataset we have acquired. We aim
1313 to undertake benchmarking experiments and harness the potential of this dataset to fine-tune state-of-the-art
1314 models, enhancing their performance and inclusivity.
1315

1320 9 CONCLUSION

1321 In conclusion, the rise of AI technologies, while revolutionary, has highlighted glaring disparities in data representation,
1322 especially for marginalized social groups such as the disability community. Our research offers an in-depth exploration
1323 of community-led data collection using StammerTalk, an online community for Chinese-speaking people who stutter,
1324 as a case study. We found that grassroots initiatives can produce authentic and high-quality datasets, driven by intrinsic
1325 motivations that foster meaningful contributions and community connections. Participants derived empowerment,
1326 personal skills, and camaraderie from the process, illustrating the huge benefits beyond the dataset output. However,
1327 challenges arise due to limited resources and the constraints of current socio-technical infrastructures, leading to
1328 complexities in navigating global geo-political tensions and data regulations. It's imperative for stakeholders – ranging
1329 from industries to academia and policymakers – to recognize and invest in building robust infrastructures that empower
1330 marginalized communities in shaping their data narratives and AI-driven experiences.
1331
1332
1333

1336 REFERENCES

- 1337 [1] Adriana Alvarado Garcia and Christopher A. Le Dantec. 2018. Quotidian Report: Grassroots Data Practices to Address Public Safety. *Proc. ACM*
1338 *Hum.-Comput. Interact.* 2, CSCW, Article 17 (nov 2018), 18 pages. <https://doi.org/10.1145/3274286>
- 1339 [2] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel
1340 White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software*
1341 *and technology*. 333–342.
- 1342 [3] Anna Bleakley, Daniel Rough, Abi Roper, Stephen Lindsay, Martin Porcheron, Minha Lee, Stuart Alan Nicholson, Benjamin R Cowan, and Leigh
1343 Clark. 2022. Exploring Smart Speaker User Experience for People Who Stammer. In *Proceedings of the 24th International ACM SIGACCESS Conference*
1344 *on Computers and Accessibility*. 1–10.
- 1345 [4] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and
1346 Richard E. Ladner. 2021. The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective. *ACM Trans. Access. Comput.* 14, 2,
1347 Article 7 (jul 2021), 45 pages. <https://doi.org/10.1145/3436996>
- 1348 [5] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. "Why would anybody do this?" Understanding Older Adults' Motivations and
1349 Challenges in Crowd Work. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2246–2257.
- 1350 [6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of*
1351 *the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo
1352 Wilson (Eds.). PMLR, 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>

- 1353 [7] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases.
1354 *Science* 356, 6334 (2017), 183–186.
- 1355 [8] Leigh Clark, Benjamin R Cowan, Abi Roper, Stephen Lindsay, and Owen Sheers. 2020. Speech diversity and speech interfaces: Considering an
1356 inclusive future through stammering. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.
- 1357 [9] Marianne Dee and Vicki L Hanson. 2014. A large user pool for accessibility research with representative users. In *Proceedings of the 16th international
1358 acm sigaccess conference on computers & accessibility*. 35–42.
- 1359 [10] Xianghua Ding, Patrick C Shih, and Ning Gu. 2017. Socially embedded work: A study of wheelchair users performing online crowd work in china.
1360 In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 642–654.
- 1361 [11] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge:
1362 Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.
- 1363 [12] Peter Howell, Stephen Davis, and Jon Bartrip. 2009. The university college london archive of stuttered speech (uclass). (2009).
- 1364 [13] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in
1365 coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331.
- 1366 [14] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Craig Denuyl. 2020. Social Biases in NLP Models
1367 as Barriers for Persons with Disabilities. In *Proceedings of ACL 2020*.
- 1368 [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.
1369 *International journal of speech-language pathology* 14, 6 (2012), 543–556.
- 1370 [16] Rie Kamikubo, Utkarsh Dwivedi, and Hernisa Kacorri. 2021. Sharing practices for datasets related to accessibility and aging. In *Proceedings of the
1371 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.
- 1372 [17] Rie Kamikubo, Kyungjun Lee, and Hernisa Kacorri. 2023. Contributing to Accessibility Datasets: Reflections on Sharing Study Data by Blind People.
1373 In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing
1374 Machinery, New York, NY, USA, Article 827, 18 pages. <https://doi.org/10.1145/3544548.3581337>
- 1375 [18] Rie Kamikubo, Lining Wang, Crystal Marte, Amnah Mahmood, and Hernisa Kacorri. 2022. Data Representativeness in Accessibility Datasets: A
1376 Meta-Analysis. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (*ASSETS '22*).
1377 Association for Computing Machinery, New York, NY, USA, Article 8, 15 pages. <https://doi.org/10.1145/3517428.3544826>
- 1378 [19] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etamad. 2021. Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning.
1379 *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2986–2999.
- 1380 [20] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023.
1381 From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. In *Proceedings of the 2023 CHI
1382 Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA,
1383 Article 361, 16 pages. <https://doi.org/10.1145/3544548.3581224>
- 1384 [21] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey Bigham. 2021. Sep-28k: A Dataset for Stuttering Event Detection from
1385 Podcasts with People Who Stutter. <https://arxiv.org/pdf/2102.12394.pdf>
- 1386 [22] Qisheng Li, Krzysztof Z. Gajos, and Katharina Reinecke. 2018. Volunteer-Based Online Studies With Older Adults and People with Disabilities. In
1387 *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) (*ASSETS '18*). Association for
1388 Computing Machinery, New York, NY, USA, 229–241. <https://doi.org/10.1145/3234695.3236360>
- 1389 [23] Qisheng Li, Josephine Lee, Christina Zhang, and Katharina Reinecke. 2021. How Online Tests Contribute to the Support System for People
1390 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>
- 1391 [24] Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q
1392 Nelson, Jordan R. Green, and Katrin Tomanek. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project
1393 Euphonia.
- 1394 [25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning.
1395 *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. <https://doi.org/10.1145/3457607>
- 1396 [26] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From
1397 People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (*FAccT
1398 '21*). Association for Computing Machinery, New York, NY, USA, 52–63. <https://doi.org/10.1145/3442188.3445870>
- 1399 [27] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of
1400 commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.
- 1401 [28] Nan Bernstein Ratner and Brian MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders* 56
1402 (2018), 69–80.
- 1403 [29] Participatory Data Stewardship. 2010. A Framework for Involving People in the Use of Data. Report.(2021, September). Ada Lovelace Institute.
- 1404 [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>

- 1405 [31] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann.
1406 2021. Disability-First Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data
1407 Collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (Virtual Event, USA) (ASSETS '21)*.
1408 Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. <https://doi.org/10.1145/3441852.3471225>
- 1409 [32] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser.
1410 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63.
- 1411 [33] Ana Rita S Valente, Luis MT Jesus, Andreia Hall, and Margaret Leahy. 2015. Event-and interval-based measurement of stuttering: a review.
1412 *International Journal of Language & Communication Disorders* 50, 1 (2015), 14–30.
- 1413 [34] Maggie Walter and Michele Suina. 2019. Indigenous data, indigenous methodologies and indigenous data sovereignty. *International Journal of*
1414 *Social Research Methodology* 22, 3 (2019), 233–243.
- 1415 [35] Janis Wong. 2023. Data Practices and Data Stewardship. *Interactions* 30, 3 (may 2023), 60–63. <https://doi.org/10.1145/3589133>
- 1416 [36] Shaomei Wu. 2023. “The World is Designed for Fluent People”: Benefits and Challenges of Videoconferencing Technologies for People Who Stutter.
1417 In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- 1418 [37] J Scott Yaruss and Robert W Quesal. 2006. Overall Assessment of the Speaker’s Experience of Stuttering (OASES): Documenting multiple outcomes
1419 in stuttering treatment. *Journal of fluency disorders* 31, 2 (2006), 90–115.
- 1420 [38] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification
1421 using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).
- 1422 [39] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments:
1423 Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI*
1424 *'18)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3173574.3173789>
- 1425
- 1426
- 1427
- 1428
- 1429
- 1430
- 1431
- 1432
- 1433
- 1434
- 1435
- 1436
- 1437
- 1438
- 1439
- 1440
- 1441
- 1442
- 1443
- 1444
- 1445
- 1446
- 1447
- 1448
- 1449
- 1450
- 1451
- 1452
- 1453
- 1454
- 1455
- 1456