

Disability-First AI Dataset Annotation: Co-designing Stuttered Speech Annotation Guidelines with People Who Stutter

Xinru Tang
Department of Informatics
University of California, Irvine
Irvine, California, USA
AImpower.org
Mountain View, California, USA
xinrut1@uci.edu

Jingjin Li
AImpower.org
Mountain View, California, USA
jingjin@aimpower.org

Shaomei Wu*
AImpower.org
Mountain View, California, USA
shaomei@gmail.com

Abstract

Despite efforts to increase the representation of disabled people in AI datasets, accessibility datasets are often annotated by crowdworkers without disability-specific expertise, leading to inconsistent or inaccurate labels. This paper examines these annotation challenges through a case study of annotating speech data from people who stutter (PWS). Given the variability of stuttering and differing views on how it manifests, annotating and transcribing stuttered speech remains difficult, even for trained professionals. Through interviews and co-design workshops with PWS and domain experts, we identify challenges in stuttered speech annotation and develop practices that integrate the lived experiences of PWS into the annotation process. Our findings highlight the value of embodied knowledge in improving dataset quality, while revealing tensions between the complexity of disability experiences and the rigidity of static labels. We conclude with implications for disability-first and multiplicity-aware approaches to data interpretation across the AI pipeline.

CCS Concepts

• **Human-centered computing** → **Empirical studies in accessibility**; **Accessibility design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

AI dataset annotation, accessibility, disability, stuttering, embodied knowledge

ACM Reference Format:

Xinru Tang, Jingjin Li, and Shaomei Wu. 2026. Disability-First AI Dataset Annotation: Co-designing Stuttered Speech Annotation Guidelines with People Who Stutter. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3772318.3790405>

*Corresponding author



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
<https://doi.org/10.1145/3772318.3790405>

1 Introduction

Recent advances in artificial intelligence (AI), such as large language models (LLMs) and other generative AI models, have relied on the use of vast amounts of human data [9]. Meanwhile, a growing body of work has documented AI models' degraded, and at times discriminatory, performance for users with disabilities¹ [30, 43, 52, 67, 92, 104, 115], and identified the lack of data representing disabled people as one of the root causes [12, 43, 64, 78, 88, 104]. In response, researchers have sourced data from disabled people to create *accessibility datasets* [26, 32, 36, 46, 63, 68, 88]. Within this movement, the notion of "disability-first datasets" was proposed to emphasize the importance of centering disabled communities' experiences and interests in both the creation and application of accessibility datasets [63, 88, 98].

Despite ongoing data collection efforts, label noise in accessibility datasets remains a persistent yet underexplored problem. One notable example is VizWiz, an accessibility dataset of photos taken by blind and low-vision (BLV) individuals for visual question answering [36]. Prior work reported over 40% disagreement between crowdworkers and in-house experts on whether a photo contained privacy or quality issues [36], and only 9% unanimous agreement among annotators on the answer to associated visual questions [35]. Similar label inconsistencies have been observed in datasets of sign languages [12] and stuttered speech [58, 92].

Two main factors may have contributed to these data labeling issues: (1) the distinctive characteristics of data sourced from disabled people (e.g. blurriness in photos taken by BLV people [36]), and (2) annotators' limited understanding of disabled people's needs and experiences [35, 63, 90, 92]. While AI data annotation tasks have become increasingly interpretive [16, 85, 86, 103], most accessibility datasets are still annotated by data workers without relevant experience or expertise in disabilities to properly interpret the data. For instance, VizWiz was annotated by sighted crowdworkers [36], who reportedly were uncertain about the kinds of visual information most useful to BLV people [90]. Similarly, Sep-28k, the largest English stuttered speech dataset sourced from stuttering related podcasts and labeled by annotators "*who received training via written descriptions, examples, and audio clips on how to best identify*

¹We use both people-first (people with disabilities) and identity-first (disabled people) language to recognize diverse naming preferences within disabled communities.

each dysfluency but were not clinicians” [58], was observed by researchers who stutter to mislabel a significant amount of natural disfluencies (e.g. fillers, pauses) as stuttering disfluencies [92]².

As an initial step toward addressing these issues, our study explores a disability-first approach to accessibility dataset annotation through a case of transcribing and annotating stuttered speech. Speech disfluencies, whether natural or stuttering-induced, are typically omitted from transcriptions in existing speech AI datasets [82], resulting in the **default erasure** of stuttering by automatic speech-to-text systems [60, 63]. Even when stuttering is annotated, inter-rater agreement remains alarmingly low. For instance, Sep-28k reported an inter-rater agreement of 0.11 for identifying speech prolongations, 0.25 for blocks, and 0.39 for determining whether a clip contains disfluencies [58]. While annotating stuttered speech is inherently challenging and subjective due to its high variability and the ambiguous boundary between stuttered and fluent speech [99, 102], people who stutter (PWS) are rarely involved to address such complexities [6, 58, 83, 102]. Building on a recent shift in clinical research that prioritizes the subjective experience of stuttering over listener perspectives [22], our study aims to incorporate the embodied knowledge and lived experiences of PWS to develop more reliable annotations of stuttered speech for AI.

More specifically, our study sets out to co-design stuttered speech annotation guidelines with PWS. Following the disability-first idea [98], we developed, evaluated, and evolved our guidelines by centering the stuttering experience and identity, rather than “co-opting” existing speech annotation for so-called “edge cases” like stuttering [74]. To do so, we crafted our co-design process into three phases: 1) *Formative studies* in which we reviewed and discussed the annotation practices of existing stuttered speech datasets with AI professionals who stutter (PWS AI professionals); 2) *Co-design sessions* in which we designed, tested, and iterated a new set of stuttered speech annotation guidelines with PWS AI professionals and speech-language pathologists (SLP) specializing in stuttering; 3) *Evaluation sessions* in which we reviewed the resulting annotations of their own speech with participants who stutter.

Our experiences show that involving PWS in data annotation is crucial for representing stuttered speech in ways that align with the lived experience of stuttering, as they bring lived, deeply embodied experiences of stuttering that fluent speakers typically do not fully understand. For instance, participants who stutter shared how they attended and leveraged non-verbal acoustic cues (e.g. change in breathing patterns) to identify and distinguish different types of stuttering. Our findings also suggest the inherent subjectivity of stuttering annotation: even PWS and SLP clinicians with extensive exposure to stuttering often perceive stuttered speech differently, and acoustic signals alone could be fundamentally limited in capturing the multitudes and complexity of stuttering.

Our study makes three main contributions to the HCI literature on AI for accessibility.

- (1) *Methodological*. Extending existing accessibility data research that has primarily focused on data collection [26, 36, 47, 48, 88, 98], we explore a disability-first approach that enables

PWS to shape the interpretation of data from the outset and throughout the annotation process.

- (2) *Artifact*. The resulting guidelines from our co-design study represent the first effort to develop PWS-centered stuttered speech annotation guidelines for speech AI development (see Appendix A). In contrast to conventional annotation guidelines that focus primarily on audible speech [58, 83], our guidelines draw on the lived experiences of PWS to highlight the importance of non-verbal signals (e.g. breathing, pauses) in communication, and caution against the common practice of trimming non-speech audio during automatic speech recognition (ASR) model training [56].
- (3) *Empirical*. Our study identifies underexplored challenges in annotating stuttered speech (see Appendix B), with broader implications for annotating accessibility datasets. As illustrated by stuttering, disability experience is inherently dynamic and situated, defying the static, categorical, and often binary framings typically imposed in AI data annotation. We therefore urge the AI research community to recognize the embodied knowledge of disabled people and to embrace greater multiplicity and interpretive flexibility when engaging with disability data throughout the AI pipeline.

Positionality. This study is shaped by our positionality as researchers working at the intersection of HCI, accessibility, machine learning, and speech technologies. Our research team includes both PWS (Wu) and non-PWS members (Tang and Li), with the PWS member serving as the senior author and guiding the research direction. Our relationships with the stuttering community are not uniform and have shaped how we approached this work. Tang entered the topic as a relative newcomer to stuttering; however, she brought prior experience in participatory and disability-centered research [31, 94, 95, 97, 101], which informed her approach to listening and collaborative decision-making. Li and Wu have long-standing involvement with the stuttering community, including efforts to develop stuttering-friendly technologies [60–64, 105, 106] and to support stuttering advocacy. Li and Wu also hold deep personal and professional ties within the community, which have influenced not only recruitment and design decisions but also our interpretation of the values, tensions, and priorities surfaced throughout this research. Collectively, our understandings of stuttering and disability are informed by our personal experiences and academic training within U.S.-based institutions, such as disability activism, critical disability studies, the neurodiversity movement, and the affirmative approach in stuttering research and therapy (e.g., [20–22]). We acknowledge the diversity of experiences and orientations within the stuttering community and have tried our best to respect such diversity throughout our study.

2 Related Work

2.1 Stuttering as a Contested Disability

Stuttering is a neurodevelopmental condition characterized by involuntary disfluencies in speech, affecting over 1% of the global population [77]. The field of SLP has historically framed stuttering as a problem to be fixed, relying on external observations and control [22, 99]. Yet, the measurement of stuttering has been a contested

²Authors of Sep-28k did not specify whether the annotators stutter or not [58]. We can only assume the annotators are not PWS as such information would have been provided if they are.

issue in SLP due to its variability and the wide range of associated behaviors [109]. Patterns and frequency of stuttering differ not only across individuals but also within the same speaker over time and in response to changing contexts, such as task, setting, or conversational partner [100, 109]. As a result, having an operationalized definition of stuttering has proven to be a difficult task in SLP, and currently there is no universally agreed-upon definition [44]. Annotating stuttered speech is found to be highly subjective task even for SLP professionals [102].

Pushing back on the medical model of stuttering, an increasing number of voices from the stuttering community have called for the inclusion of PWS in shaping knowledge about stuttering [22, 99] as well as the prioritization of PWS’s perspectives and lived experience in stuttering research and therapy [19, 21, 22, 99]. Constantino contends that spontaneity, rather than fluency, should be the guiding objective of speech therapy [22]. Other research further emphasizes the affective, behavioral, and cognitive dimensions that shape the lived experience and intimate understanding of stuttering [99]. Our work contributes to these ongoing efforts by centering the lived experiences of PWS in data annotation as a central way of shaping speech AI technologies.

2.2 Disability-First AI Dataset Creation

Theodorou et al. proposed *disability-first* as an approach “*serving a disability community first*,” which “*stands in opposition to mainstream ML datasets and approaches which are later augmented or co-opted to address issues of importance to disabled communities*” [98]. Reflecting similar ideas and in response to general-purpose datasets that often fail to adequately represent or prioritize disabled people’s perspectives and needs [18, 64, 92], researchers have created disability-first datasets such as VizWiz [36], Orbit [68], ASL Citizen [26], and AS-70 [32, 63] – all of which directly sourced from disabled communities. Other work has explored collection mechanisms aimed at supporting disabled people as data contributors, including accessible designs of privacy access control for blind users [48], community-led data collection method for stuttered speech [63], and data anonymization tools designed to protect deaf³ users in sharing sign language videos [13]. However, these efforts face ongoing challenges, including open questions around fair compensation, consent, and ethical engagement with disabled communities [78].

Despite recent disability-first data collection efforts, a critical gap remains in data annotations. Disabled people are typically treated solely as the source of data during data collection, excluded from other crucial stages of the AI development pipeline such as data annotation and model design [61]. Annotation of accessibility datasets often relies on crowd workers who lack familiarity with the target communities. VizWiz, one of the largest and most widely used visual accessibility datasets sourced from BLV people, relies on sighted crowd workers to provide labels and determine label accuracy through inter-annotator agreement [36]. However,

sighted crowd workers often apply inconsistent interpretive standards shaped by their own assumptions and biases [90]. These inconsistencies can significantly impact benchmarking and the usability of downstream applications for BLV users [31, 50]. Similar limitations are common in accessibility datasets, such as sign languages [12, 27] and stuttered speech [58], which rarely involve affected communities in the annotation process.

These issues in accessibility dataset annotation align with broader calls in critical AI research to shift attention to the underlying systems of power in AI development [5, 70]. Prior research has shown that so-called ‘ground truth’ labels are shaped by specific contexts, such as annotators’ backgrounds and expertise, as well as broader organizational control [1, 76]. For example, studies of industry-scale data collection revealed that the pursuit of label precision is often driven by market-oriented values such as objectivity, standardization, and corporate interests [49, 51, 71, 112, 113]. Other research has uncovered the ambiguity and uncertainty inherent in socially constructed labels, such as gender [85], accent [81], and content toxicity [1, 16]. Consequently, a growing body of scholarship emphasizes the need to document the production processes and contextual factors involved in dataset creation [72, 73]. Research has also called for more reflective data practices, advocating for the treatment of bias or error as a site of negotiation rather than simply as model failures [16, 65]. Our work extends this line of research to the domain of accessibility and disability through a case study of stuttered speech annotation.

2.3 Stuttered Speech Datasets

Existing stuttered speech datasets have been developed primarily for two purposes: corporate initiatives and academic research. Much recent development has been driven by corporate interest in ASR and a growing recognition of the need for more diverse speech data [57, 58, 66, 114]. However, access to these datasets is often restricted, as they typically involve sensitive or proprietary data, except when sourced from public domains such as podcasts [58]. As a result, current research has mainly relied on publicly available datasets collected for academic use.

Datasets collected by academia primarily come from two fields: SLP and machine learning (see Table 1 for an overview). As training materials for speech pathologists, SLP datasets are typically limited in size. In contrast, datasets from the ML community tend to be more scalable, but still limited by annotation granularity and representativeness. For instance, as the largest publicly available English stuttering dataset, Sep-28k does not have transcriptions and provides only clip-level annotations rather than word-/syllable-level labels [58]. LibriStutter, another influential English stuttering dataset, did not collect speech from PWS but injected synthetic stuttering sounds into read speech from fluent speakers [53].

An underlying concern across these efforts is the limited power that PWS have in the creation of stuttered speech datasets. While PWS have frequently contributed as data sources, their influence on other critical aspects such as annotation frameworks, labeling guidelines, and the interpretation of stuttering behaviors remains minimal. For example, Sep-28k consists of over 28k three-second audio clips curated from public podcasts featuring PWS; however, the annotators’ backgrounds remain unclear. Among the five stuttering

³Deaf communities in the U.S. often capitalize ‘D’ in ‘Deaf’ to emphasize a shared cultural identity. We do not differentiate between Deaf and deaf in our writing, as this distinction is increasingly contested within deaf studies [54]. We use deaf to acknowledge the fluidity of identity and to recognize that access to deaf cultural resources itself can be a form of privilege. We use deaf and hard-of-hearing (DHH) to encompass broader populations who have hearing disabilities.

| Dataset | Field | Speakers | Tasks | Transcripts | Annotations | Annotators | Languages |
|-------------------|-------|-----------------|--|-------------|----------------|---|--|
| FluencyBank* [83] | SLP | 22 CWS & 38 AWS | conversation, reading | Yes | N/A | N/A | English |
| UCLASS* [42] | SLP | 25 CWS | conversation | Yes | N/A | N/A | English |
| KSoF [7] | SLP | 37 PWS | reading, spontaneous speech (e.g., dialogues, scene description) | No | clip-level | non-PWS students | German |
| LibriStutter [53] | ML | 50 non-PWS** | audiobook | Yes | clip-level | not reported | English |
| AS-70 [32] | ML | 72 AWS | conversation, voice commands | Yes | word-level | non-PWS professional annotators; reviewed by non-PWS supervisors and a PWS researcher | Mandarin |
| Boli [6] | ML | 28 PWS | reading, scene description (spontaneous) | Yes | syllable-level | not reported | English, Hindi, Telugu, Bengali, Marathi, Assamese |
| Sep-28k [58] | ML | PWS podcasts | podcast | No | clip-level | trained non-PWS | English |
| Sep-28k-SW [92] | ML | PWS podcasts | podcast | Yes | syllable-level | PWS researchers | English |

* Limited to the transcribed portion of the dataset.

** Synthetic stutters were injected into fluent speech.

Abbreviations: AWS – adults who stutter; CWS – children who stutter; PWS – people who stutter.

Table 1: Examples of major stuttered speech datasets developed in SLP and ML research.

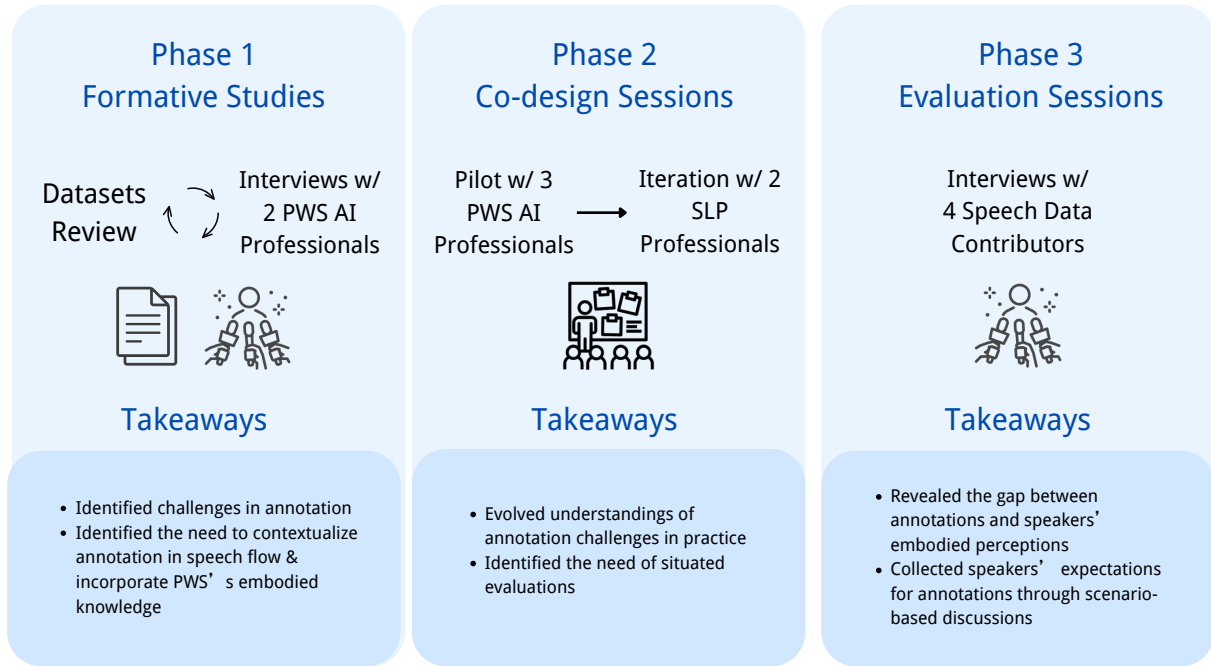


Figure 1: Study flow with main takeaways at each stage.

labels it adopts, inter-annotator agreement scores were moderate for word repetitions (0.62) and interjections (0.57), but dropped significantly for sound repetitions (0.40), no disfluencies (0.39), blocks (0.25), and prolongations (0.11) [58]. Furthermore, a recent study conducted by a team of PWS researchers (Sridhar and Wu) observed

a significant number of “mistakes” in the original event labels and adjusted the labeling for over 25% of the 2,621 clips in their sample for their benchmarking studies [92]. Their work calls into question the original labels of Sep-28k and yields a re-annotated subset, which we hereafter refer to as Sep-28k-SW. An exception to these

efforts is AS-70, where the quality control team included one PWS member who reviewed the annotations. However, AS-70 focused on Mandarin and the original annotations were still generated by non-PWS crowdworkers. Our work extends this line of research to English and centers PWS perspectives from the outset.

3 Study Overview

This study is part of a larger, ongoing initiative to develop representative speech datasets with and for PWS to make speech AI models more inclusive of stuttering. Prior to the present study, we have recorded conversational and reading speech from 51 English-speaking PWS, following a disability-first process similar to what was described in [63]. The present study was directly motivated by labeling issues identified in existing stuttered speech datasets [57, 92]. As an alternative to conventional annotation approaches, our work explores the development of PWS-led guidelines to more accurately and authentically represent stuttering in transcription and annotation practices. Our study consists of three stages, as shown in Figure 1.

- (1) *Phase 1: Formative Studies.* We first reviewed and analyzed existing annotated datasets to identify gaps and limitations. Specifically, we examined different annotated versions of the Sep-28k dataset [58, 92] and interviewed two PWS AI professionals experienced in stuttered speech annotation to gather feedback on current practices and opportunities for improvement. We focused on Sep-28k due to its widespread use⁴, its public availability of annotations produced by PWS [92] and non-PWS [58], and the reported discrepancy between these two versions [92]. At the conclusion of this phase, we drafted preliminary annotation guidelines to serve as the foundation for subsequent co-design sessions.
- (2) *Phase 2: Co-design.* We refined, piloted, and iterated our guidelines with PWS AI professionals and SLPs specializing in stuttering. Together, we applied the preliminary guidelines to annotate stuttered speech samples, reflected on our experience and challenges in the application, and incorporated PWS participants’ embodied knowledge of stuttering into the guidelines to mitigate challenges and better represent stuttering experiences.
- (3) *Phase 3: Evaluation.* We evaluated the refined guidelines by reviewing the annotation of participants’ own speech with four PWS data contributors.

All study sessions were conducted over Zoom with participants’ consent to record for analysis. As all but one participant were PWS (see Table 3), we adopted proactive strategies to create a supportive communication environment, including confirmation of turn taking, reserving buffer time to reduce time pressure, and encouraging the use of non-verbal channels such as emojis, body gestures, chat, and virtual whiteboard. Recognizing that listening to one’s own stuttered speech can be uncomfortable [105], we ensured participants were fully informed and explicitly consented before playing the recordings of their speech during the evaluation phase.

This study is exploratory and interpretative in nature. Our goal is to examine how the lived experiences of stuttering could inform and

| Stuttering Events | Definitions |
|------------------------|---|
| prolongations | Elongated syllable: “M[mmm]ommy” |
| block | Gasps for air or stuttered pauses |
| sound repetition | Repeated syllables: “I [pr-pr-pr-]prepared dinner” |
| word/phrase repetition | Repeated words: “I made [made] dinner” |
| interjection | Filler words, e.g., “um”, “you know” |

Table 2: Definitions of stuttering events in Sep-28k [58].

improve the annotation of stuttered speech. Therefore, aside from the descriptive statistical analysis conducted during our review of stuttered speech datasets in the formative stage, we employed reflexive thematic analysis as our primary method [14]. The data used in analysis involved notes from dataset reviews, transcripts from interviews, and video recordings of co-design workshops, all collected with participants’ informed consent. Our analysis includes iterative and ongoing theme development along with data collection based on patterns of shared meaning among the data [14]. We followed the same approach for all our qualitative data analysis. The lead author conducted open inductive coding of the cumulative data and regularly discussed the themes within the research team. The discussions centered on the distinct perspectives participants contributed to the transcription and annotation of stuttered speech for AI, as well as the challenges and trade-offs involved. We iterated on the coding process until reaching consensus over all the themes. This sole-coder approach aligns with the interpretative nature of qualitative research, and senior researchers were involved at all stages to enhance the reliability of the findings [17].

4 Formative Studies

We began with formative studies consisting of two parts: (1) an iterative analysis of Sep-28k [58] and Sep-28k-SW [92]; and (2) interviews with two PWS AI professionals who had annotated stuttered speech for their work.

4.1 Methods

4.1.1 Datasets Review. To explore annotation inconsistencies, we compared Sep-28k (the original version produced by crowdworkers [58]), and Sep-28k-SW (a re-annotated subset by Sridhar and Wu, both of whom self-identified as PWS [92]). The audio recordings in Sep-28k are broken into 3-second clips and annotated by three annotators with five binary labels: word repetition, sound repetition, block, prolongation, and interjection (see Table 2 for definitions) [58]. Sep-28k-SW was developed based on Sep-28k to benchmark ASR performance for stuttered speech [92]. Sridhar re-annotated a subset of Sep-28k, prioritizing clips with unanimous annotator agreements in the original Sep-28k. They reported making adjustments to over 25% of the labels – mostly due to the confusion between natural and stuttering disfluencies, and called for “close attention to the content, flow, and voice quality” during stuttering annotation [92]. The lead author reviewed both datasets between March and June 2025. She identified and listened to audio

⁴Over 170 citations since its release in 2021 according to Google scholar (<https://scholar.google.com/scholar?cites=10057318411144677061>, accessed on 1/27/2026).

clips with different labels⁵ between Sep-28k and Sep-28k-SW, and documented her thoughts on the difference alongside each clip for further analysis.

4.1.2 Interviews. Alongside our dataset review, we interviewed two PWS AI professionals (Charan and Rong in Table 3) in March 2025 to incorporate their perspectives as PWS speakers and expertise in stuttered speech annotation. Besides being PWS, both have worked with English stuttered speech datasets and had extensive experience annotating stuttered speech. We recruited them from our personal network and received their informed consent to conduct audio/video recorded interviews for this study.

The interviews were semi-structured and guided by an interview protocol including questions about their understanding of stuttering, experiences with annotation work, etc. (see Appendix C). Both interviews were held over Zoom and each lasted about one hour. To facilitate reflection, we selected one stuttered speech audio sample in each participant’s native language from two publicly available stuttered speech datasets [32, 58] and invited them to explain how they considered the transcriptions and interpreted the stuttering event labels. We transcribed the interview audio recordings for analysis. The analysis resulted in four themes on challenges in annotations as presented in 4.2.

4.2 Findings

In line with Sridhar and Wu’s findings [92], our analysis reveals substantial annotation disagreements between Sep-28k and Sep-28k-SW. Table 4 shows the distribution of annotation disagreements across five types of stuttering events. While disagreements occur across all stuttering types, Sep-28k consistently labels more stuttering events than Sep-28k-SW – resulting in more *false positives* than *false negatives* in Table 4. This pattern highlights a significant gap in how stuttering is perceived and understood by PWS and non-PWS annotators. Our fine-grained dataset reviews and interviews with PWS AI professionals allowed us to identify four key factors that may contribute to this gap.

4.2.1 Lack of Embodied Understandings of Stuttering. A major source of the difference arises from a limited understanding of stuttering by non-PWS. Both interview participants emphasized the importance of a contextual and embodied approach to annotation. Charan observed that a significant portion of the “errors” in Sep-28k’s labels stemmed from labeling “*clearly fluent speech*” as stuttered:

“I think the labels [of Sep-28k] just kind of followed a rigid playbook. They didn’t use any intuition. They had like a set of rules: If they hear ‘um,’ it’s an interjection. If a sound is dragged, it’s a prolongation. If there’s some pause, it’s a block.”

Rather than solely focusing on overt speech signals, he emphasized that stuttering should be perceived as situated within the flow of speech, noting that speakers may sometimes naturally prolong a word for non-verbal expression rather than due to stuttering:

“I feel like I can tell that ‘Oh, their voice is strained. That’s stutter’ versus ‘It’s answering a question and there’s a long dragged out ‘um.’ That’s not stutter. It’s just them thinking.”

Echoing this quote, we observed many instances in Sep-28k where clips with natural disfluencies – such as intentional interjections, prolonged interjection (e.g., *ummmmm* to agree), pause because of thinking – were labeled as stuttering.

Participants described how they drew on their embodied knowledge of stuttering when interpreting the data. Charan mentioned **accessory signals** such as “*breathing, tone of speech, and speed of talking*”. Similarly, Rong explained the **kinematic signals** he used to identify blocks, derived from his own sensation of stuttering: “*You can hear the person is trying hard when they pause. Their teeth or vocal cords might produce small sounds.*” In another instance, when discussing a case that the lead author found difficult to distinguish between prolongation and block, Rong explained that prolongation “*should be continuous in the flow when pronouncing a sound*,” which is distinct from the choppy sounds associated with blocks.

Moreover, both participants noted the difficult case when PWS try to circumvent overt stuttering, yet they could better notice these “tricks” for masking stutters:

“If I practice properly, I can hide it [stuttering] pretty well. But I’ll know where there was a small catch where no one else would notice.” (Charan)

“I can feel when they tried to replace the word. For example, they might pause at a place where non-stutterers typically wouldn’t stop. They might speak fast but drag out the last word before he got stuck.” (Rong)

While these less observable signals might feel intuitive to them, both emphasized that these perceptions are hard to translate into prescriptive rules. As Charan noted,

“The way I’ve focused on the breathing is very subjective. I couldn’t describe a way to quantify breathing. Measures like speed of talking, the words per minute, are not a great way.”

These subjectivities and the reliance on embodied knowledge make perceiving stuttering challenging for those without lived experience. The difficulty is further compounded when annotators must codify the subtle and fluid audio cues into written transcription.

4.2.2 Challenges in Transcribing Disfluent Speech Verbatim. The option to transcribe stuttered speech as it is is important to PWS [63], but significantly complicated the annotation process. A notable challenge lies in distinguishing between sound-level and word-level repetitions, particularly with words that only have one syllable such as *add*, *how*, and *what*, and when speakers speak fast. Rong highlighted the challenges of converting and preserving speech disfluencies into text transcriptions,

“You might need to listen to the clips many many times. You have to count the words they repeated. Sometimes they repeat very fast. In such cases, it’s hard to get an accurate transcription and annotation even if you keep listening.”

⁵As each clip has only one single label in Sep-28k-SW but three labels in Sep-28k by three different annotators, we determined the final labels for Sep-28k using majority voting.

| Participant | Phase | Background | Community Activity | Relevant Experiences |
|-------------|-------------------------|----------------------------------|--|---|
| Charan | Formative; Co-design | PWS, student, AI researcher | Host of stuttering self-help groups; volunteer at stuttering support organizations. | Used and annotated Sep-28k for research [92]. Co-author of Sep-28k-SW dataset. |
| Rong | Formative; Co-design | PWS, tech worker in AI fields | Co-founder and executive director of StammerTalk community. | Used Sep-28k for work. Co-creator of Mandarin Stuttered Speech dataset [63] (designed annotation guidelines; revised annotations by non-PWS). Published authors on speech AI for stuttered speech [32, 60]. |
| Jia | Co-design | PWS, SLP professional | Organizer and host of multiple stuttering self-help groups; co-founder, director, and/or advisor for multiple stuttering community organizations | Annotated stuttered speech in educational and clinical settings |
| Kerrigan | Co-design | non-PWS, SLP professional | Director and programming associate for multiple stuttering community organizations | Annotated stuttered speech in educational and clinical settings |
| Adedotun | Evaluation | PWS, medical student | Volunteer at stuttering support organizations | Contributed speech to dataset |
| Amina | Evaluation | PWS, PhD student | | Contributed speech to dataset; prior experience transcribing own speech |
| Tatianna | Evaluation | PWS, small business owner | | Contributed speech to dataset |
| Benji | Evaluation | PWS, tech worker in AI fields | Volunteer at stuttering support organizations | Contributed speech to dataset; used multiple stuttered speech datasets for personal projects; deep technical expertise in speech AI. |

Table 3: Participants’ backgrounds. All of our participants in the Formative and Co-design phases have extensive experience organizing and leading stuttering community activities and advocacy efforts. All participants in the Evaluation Phase have previously contributed speech data and were asked to evaluate the annotation of their own speech. All participants have given explicit permission to use their real names.

| "Does the clip contain ...?" | FN | FP | total | % in Sep-28k-SW |
|------------------------------|-----|-----|-------|-----------------|
| prolongation | 35 | 288 | 323 | 12.32% |
| block | 67 | 251 | 318 | 12.13% |
| sound repetition | 110 | 119 | 229 | 8.74% |
| word repetition | 55 | 88 | 143 | 5.46% |
| interjection | 85 | 193 | 278 | 10.61% |

Table 4: Distribution of annotation disagreements between Sep-28k [58] and Sep-28k-SW [92]. Treating the binary labels in Sep-28k-SW as *actual* labels and the ones in Sep-28k as *predictions*, FN (false negative) shows the number of clips on Sep-28k-SW labeled as 0 (absent) in Sep-28k but 1 (present) in Sep-28k-SW, and FP (false positive) are clips labeled as 1 (present) in Sep-28k but 0 (absent) in Sep-28k-SW. The percentage is calculated over total number of clips in Sep-28k-SW (2,621). Each clip can have multiple labels.

4.2.3 Inherent Subjectivity in Speech Perception. In some cases, participants recognized that speech perception is inherently subjective. Drawing from his experience training non-PWS annotators, Rong found that while annotations could be consistent in “roughly 80% of cases”, achieving complete agreement across all instances is “impossible”. As he said, “*The commonsense to everyone is different... It’s impossible to make everyone the same, even when you annotate fluent speech.*” Echoing this quote, we observed many annotation

differences in prolongations, rooted in annotators’ subjective judgments of what prolongations sound like, particularly with vowel sounds in short words such as *hear*, *been*, *ends*, *oh*, *two*, *so*, and *okay*.

4.2.4 Distortions Introduced by Audio Segmentation. All of the above factors were further influenced by how the audio was segmented. As Sep-28k divides audio into strictly 3-second clips [58], it is often difficult to determine stuttering events without access to the conversational flow and context. For example, segmentation can obfuscate whether a sound is part of a word or a disfluency caused by a block. In some cases, it is difficult to distinguish between words due to the cut, for instance, *um* versus *I’m*, or *and* versus *uh*. Charan thus suggested “*stringing together the clips instead of having 3-second clips.*”

5 Co-Design Sessions

Drawing on insights from our formative studies, we co-designed a set of annotation guidelines with PWS AI professionals and SLP professionals through a series of remote workshops conducted via Zoom between June and July 2025. To facilitate participants’ thinking and discussions, we started with preliminary guidelines derived from our formative studies and worked with participants to test, critique, and iterate the guidelines through co-design sessions. Informed by our formative studies, we guided these sessions with three desirable qualities for annotations: (1) contextualized in

speakers' speech flow; (2) prioritizing the PWS's embodied knowledge; (3) reflective of the trade-offs introduced by subjective speech perception.

We began with PWS sessions to anchor in PWS's perspectives, followed by SLP sessions to integrate their professional perspectives. Overall, our co-design sessions sought to address two key questions when annotating stuttered speech:

- **Ontology:** what labels should be included to represent stuttering events?
- **Consistency:** where are annotation disagreements most likely to occur, and how to resolve them?

5.1 Methods

5.1.1 Preliminary Guidelines. We developed a set of preliminary guidelines to serve as a starting point for supporting and guiding discussions during our co-design sessions (available in the supplementary material). These guidelines were designed based on insights from our formative studies, co-author Wu's lived experience with stuttering, and a review of existing annotation guidelines and practices for stuttered speech, including FluencyBank [83], Sep-28k [58], and AS-70 [32]. The preliminary guidelines consisted of two components: (1) **an annotation framework for stuttering**, including the list and definition of stuttering events to be labeled, and (2) **general transcription practices**, such as how to transcribe accents and dialects, acronyms, and sensitive information. We decided it was crucial for the guidelines to support both verbatim and semantic transcriptions, as well as stuttering event detection, as these use cases are shown to be meaningful to the stuttering community [62, 64] and the development of disfluency friendly technologies [32, 92].

We adopted the five stuttering events used in Sep-28k as our initial ontology, as these events represent the core disfluency types identified across prior work [32, 58, 63, 83]. We used the following markup for the five event types: */r*: word or phase repetition; */s*: sound repetition; */b*: blocks; */p*: prolongation; */i*: interjection. The use of event markup allowed us to embed fine-grained stuttering event annotations within the text transcriptions. We also adopted elements from other annotation frameworks, such as FluencyBank [83], to address the limits of Sep-28k (e.g. the lack of guidelines for speech transcriptions). The notations were designed to accommodate different downstream applications. For instance, we used brackets to indicate repeated segments, e.g., *[pr-pr-pr]/sprepare* (three repetitions of the *pr* sound). Leaving the final *"pr"* outside the brackets makes it easier to produce semantic transcriptions by removing all notations (e.g. *"/s"*) and bracketed content, while stripping the symbols (notations and *[]*) would generate verbatim transcriptions. This scheme allows flexibility for different use cases. Finally, we reviewed established annotation practices for both stuttered and fluent speech [32, 45, 58, 63], incorporating relevant components such as transcribing speech with accents and dialects, handling sensitive information, and representing numbers, symbols, and acronyms.

5.1.2 Workshop Sessions with PWS AI Professionals. We first conducted workshops with two PWS AI Professionals to center PWS's perspectives (Charan and Rong in Table 3). We prioritized the perspectives of PWS in our discussions, while non-PWS members of

our research team primarily served as observers and participated in annotation to support these discussions from non-PWS perspectives. We held three iterative sessions, each lasting between 40 and 60 minutes. Rong participated in one session, while Charan completed all activities.

Each session included the following activities: (1) First, PWS participants discussed the appropriateness and clarity of the annotation framework for stuttering in our preliminary guidelines, in comparison to other existing stuttering annotation frameworks.. These discussions helped to refine and familiarize the participants with our annotation framework for subsequent annotation activities. (2) Second, using the annotation framework, we annotated a 5-minute audio clip of stuttered speech we had collected. (3) Third, we compared our annotations and discussed the differences. We documented our insights (e.g., heuristics for identifying stuttering events) and iterated our guideline before the next session. We recorded videos of all sessions and compiled the artifacts developed, including document histories, for qualitative analysis. This analysis yielded themes on participants' perceptions of the labeling framework, their suggestions, strategies for resolving differences, and views on annotation challenges.

5.1.3 Workshop Sessions with SLP Professionals. Following the workshops with PWS, we invited two SLP professionals (Jia/PWS, Kerrigan/non-PWS) with speech annotation experiences to further test and iterate on the guidelines. See Table 3 for their backgrounds. The perspectives from SLP professionals are valuable, as they possess both clinical knowledge of stuttering and extensive experience in evaluating and annotating stuttered speech. As such, their participation could not only help assess the performance of our guideline but also enrich it with insights from a professional perspective. Jia and Kerrigan were also chosen for this study as they had participated in stuttering advocacy and were well aware of the values and goals of the stuttering community.

Given the time and attention required for annotation, we asked participants to complete their annotations prior to the session. We shared with them two audio clips of stuttered speech from our inclusive speech AI project (one 10-minute and one 5-minute clip), drawn from two participants who reported severe and moderate stuttering. We deliberately chose these samples to probe how the guidelines handle different levels of complexity in stuttered speech. To simulate the annotation process, we asked them to use Praat, a widely used speech annotation tool in SLP field [11] and for creating AI speech datasets [32]. We imported machine (Otter.ai) generated transcripts into Praat as the starting point for annotation. Two researchers in our team reviewed the annotations by Jia and Kerrigan and highlighted the differences before the workshop. During the workshop session, we discussed these differences and gathered suggestions to refine our guidelines. The whole session lasted around 60 minutes. We followed the same analysis approach as the sessions with PWS AI professionals. The analysis resulted in themes on issues participants pointed out, suggestions they provided, and their perspectives on the challenges in annotations.

5.2 Findings from PWS Sessions

5.2.1 Labeling Frameworks. The PWS participants first discussed the labeling frameworks drawing on their own experiences with

| Example | Session | Annotation 1 | Annotation 2 | Sources of Differences |
|---------|---------|-------------------------------------|--------------------------------|--|
| #1 | PWS | [A-A-]/sAdd | [A-a-]/s[add]/radd | number of repetitions; one-syllable words |
| #2 | PWS | [how]/r H/p/bow | H/b How | word repetition vs. backtracking induced by block; prolongation vs. making sounds when encountering blocks |
| #3 | PWS | sh/bopping | [sh/p]/s shopping | embodied conceptualization vs. labeling based on audio cues |
| #4 | PWS | /b[O-O-]/sOpen | [O-]/sOpen | a light block before “open”; number of repetitions |
| #5 | SLP | [Ha-ha-ha-ha-ha-ha-]/shackathon | [ha-]/sha/p[ha-ha-ha-]/skathon | number of repetitions; mixed stuttering events |
| #6 | SLP | fr[o-o-o-o-]/som | fro/pm | varied representations of sound repetition infused with elongation |
| #7 | SLP | w/porking | working | subjectivity in perception of prolongation |

Table 5: Examples of differences in annotations surfaced in the sessions and their causes. Note that this list is not intended to be exhaustive but present typical patterns of differences.

stuttering. For example, they questioned the necessity of FluencyBank’s distinction between pauses within words and blocks before words, as both PWS participants perceived both as essentially the same experience. Two PWS participants also questioned FluencyBank’s categorization of fillers (e.g., *um*, *you know*) as *typical* disfluencies (i.e. not stuttering), noting that such interjections are often part of their stuttering experience. In general, the participants agreed with the stuttering event categorizations presented in the preliminary guidelines, and did not see the need to add or remove any labels from the framework.

Participants also discussed the value of labels and what information to encode in relation to potential use cases. In many cases, they weighed annotation granularity against the labeling workload and the technical feasibility of data processing. For example, while Rong noted the annotation of pause duration used in FluencyBank can be valuable to avoid speech models cutting PWS off, he added that this information can be inferred through algorithms if timestamps and block labels are available. Similarly, both PWS participants chose not to label typical disfluencies to reduce annotators’ workload and prioritize tasks supporting stuttered speech. In the end, both participants agreed to adopt the five stuttering events used in Sep-28k as they offered a good balance between granularity and usefulness.

5.2.2 Differences in Annotations and Resolving Strategies. The annotation differences emerged during the sessions reflected challenges observed in our formative studies, e.g., difficulties in capturing numbers of fast repetitions, and identifying subtle or nuanced stuttering events. Table 5 presents examples of major differences we identified. Around these challenges, our discussion has led to three core principles for attending to potential differences in annotation: (1) centering PWS’s embodied understanding; (2) grounding decisions in community interests; and (3) acknowledging subjectivity in speech perception.

Centering PWS’s embodied understanding. A key way participants resolved differences was by sharing their embodied experiences of stuttering, as many disagreements stemmed from the gap between living with stuttering and observing it from the outside. Consider Example #2 in Table 5 as an example. In this instance, the speaker appears to repeat the word “how” and elongate “H” in pronouncing the second “how”. However, Charan did not label it as sound repetition or elongation, basing his judgment on his embodied experiences with stuttering. As he explained,

“I think it could be easily mistaken for a prolongation, because he was trying harder and he’s still making sounds in the block. But I would still call that a block because he’s not forced to drag it out. It’s like he’s trying a door multiple times. He’s not pushing open the door slowly, like, a really heavy door that you’d want to push it open. A block, it’s like a jam door, you have to, bump into it again and again.”

In this instance, Charan based his interpretation on his embodied understanding instead of audio cues. He distinguished between intentional repetition and involuntary repetition caused by stuttering. In the context of stuttering, the former is known as backtracking [40], a technique to pause and repeat a sound, word, or syllable to avoid an experienced block or disfluency. While the door metaphor Charan used vividly illustrated this distinction, this nuanced yet fundamental difference is hard to get without extensive exposure to stuttered speech or lived experience.

Grounding decisions in community interests. Participants also discussed the trade-offs involved in obtaining accurate annotations, given the significant human labor required. These discussions were grounded in their understanding of community interests. For example, when considering the challenges of counting repetitions, Rong emphasized the importance of taking into account the intended use of these annotations:

“It depends on what the models are trained for. You only need to get the event types right if your goal is to do stuttering event detection. However, if you wanna do transcription, you might need to get the number of repetitions right as well.”

These discussions open up considerations of downstream applications and how annotation expectations may vary across contexts.

Acknowledging subjectivity in speech perception. In many cases, participants realized that speech perception is subjective in nature. A typical scenario is how to handle the omission of sounds in single-syllable words in speech flow. As Li (non-PWS co-author) observed,

“I feel we are consistent in what we heard. We all heard that there is a block before ‘is,’ but it’s hard to say it’s the word ‘is’ repeated or the sound ‘i’ repeated.”

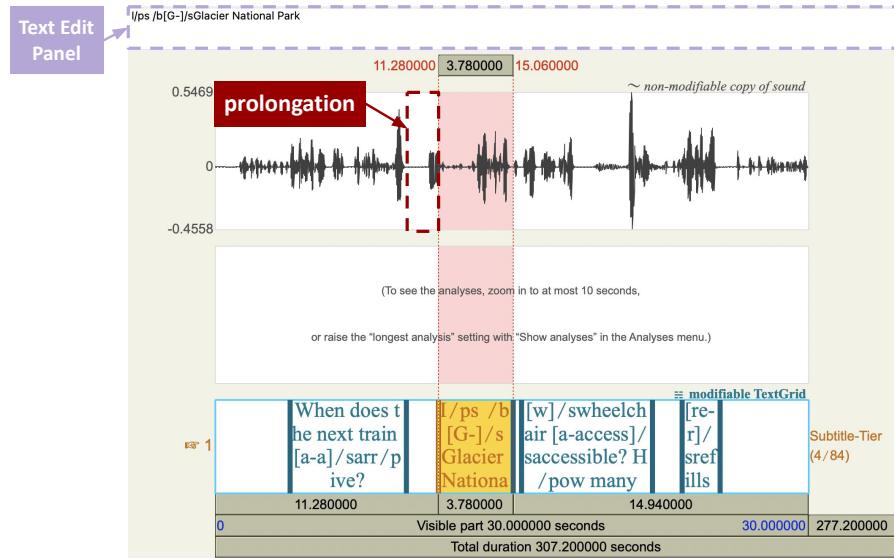


Figure 2: Screenshot of Praat interface, with a selected audio segment (in red) and corresponding editable text panel. The example shows how a prolongation might easily be overlooked due to selective listening and editing. The audio waveform corresponding to the prolongation labeled in *I/ps* occurs outside the selected segment.

In some cases, subjectivity was also evident in how stuttering was perceived across PWS, particularly in cases involving light or brief stuttering events. Consider the following discussion between Charan and Wu (PWS co-author) about a light block, for example:

- Wu: It's hard for me to hear the block after the repetition.
 Charan: I just feel there's an unnecessary pause during the sound repetition. It's not as rapid as sound repetition. There's like a pause there.
 - They repeated the audio and listened to the segment again -
 Wu: I just heard a lot of repetition of S. It's very hard to hear it [the block].

In this instance, the perception of a pause as 'unnecessary' or 'unnatural' is inherently subjective and difficult to articulate, often influenced by listeners' exposure to the speaker, embodied experiences, and judgments. These uncertain cases make an accurate representation of stuttering a nearly impossible goal to achieve.

Summary. The sessions with PWS AI professionals provided rich insights that helped us refine our guidelines. The PWS rows in Table 6 summarize changes across the sessions. A big takeaway from the discussions is the value of embodied knowledge that PWS bring to data interpretation. In response, we articulated and included key accessory and kinematic signals of stuttering participants shared, such as breathing patterns, speech flows, and vocal tension. However, the discussion also left many open questions, such as how to address the difficulties of converting speech flow into text and the inherent subjectivity with speech perception. To further explore

the challenges in practice, we shared and tested these extended guidelines with SLP professionals in subsequent sessions.

5.3 Findings from SLP Sessions

The SLP professionals brought deep expertise grounded in their extensive exposure to stuttered speech and formal training. Their practice with our guidelines confirmed the challenges identified in earlier sessions, while also surfacing two additional ones that enhanced our understanding of the subjectivity in annotations: (1) the influence of audio processing within annotation software, and (2) the challenge of representing diverse stuttering patterns. They also offered suggestions to improve our guidelines and annotation process.

5.3.1 Impact of Audio Processing Methods. A key issue emerged concerned the effect of audio segmentation and processing in Praat, the annotation software widely adopted for clinical speech analysis as well as AI speech data annotation. In Praat, users select an audio segment based on timestamps and edit annotations in the text panel. While this design makes it convenient to listen to a segment repeatedly, the segmentation could influence how the speech was perceived. Even when we avoided breaking a word or a sentence, the timestamps provided by machine transcription services could still skip over subtle blocks or prolongations occurring just before or after a word or sentence as ASR models typically neglect non-speech content [56] (see Figure 2 for an example). As Kerrigan explained,

"I'm just paying attention to the segments that have speech attached to them, and you need to be careful, because there will be segments that are not denoted as having speech in them. Then you go and you press

that segment and you play it, and it's a block, or it's a prolongation. But it's just very, very quiet, and the mic is barely picking up on it."

Building on this point, Jia noted the variability in speech segmentation can affect how much context available to speech annotators and models to interpret stuttering. Even when annotators manually segment the entire conversation, the segmentation can still vary. Jia explained, as part of the SLP training, they have learned multiple standards for speech segmentation, and could apply them fluidly in different situations. Jia's point on segmentation variability sparked discussions on the potential impact of speech segmentation methods on models' ability to process stuttered speech. Since ASR models have increasingly relied on language models to produce most probable transcripts [82], how the speech is segmented would significantly influence the amount of context available to language models.

5.3.2 Diverse and Intersecting Stuttering Patterns. Another challenge lies in how to represent the multitudes of stuttering patterns. A common scenario involves labeling hybrid stuttering patterns. Although we categorized five stuttering event types, PWS often exhibit hybrid stuttering patterns, in which multiple types of stuttering occur simultaneously or overlap. Consider Example #6 in Table 5 as an example. In this instance, while Jia labeled it as a prolongation, a more accurate way to describe what she heard was "*prolongation infused with repetition*." To her, how to represent these mixed and nuanced types of stuttering present a challenge when following a rigid guideline.

"I do think sometimes even me being a person who stutters and hearing the stuttered speech, it's kind of hard to tell, if a block is infused with repetition, right? And sometimes it blocks with prolongation... Sometimes I don't want to over identify stuttering, but at the same time I don't want to not annotate them."

The quote above shows that even with lived experiences of stuttering, the judgment can still be rather subjective because of the heterogeneity of stuttering and the overlapping nature of stuttering events. As a result, both participants agreed that there is no single correct way to represent the speech but different perspectives. Both participants emphasized that differences in annotation should not be necessarily viewed as mistakes but as perspectives that can "*enrich each coder's perspectives*", as Kerrigan explained. Jia agreed, "*we both did a good job, and we identified where the stuttered speech is, and it's important we're pretty consistent within ourselves. I think that's the best you can get from a coder.*"

Although they annotated the speech differently, both believed that their approaches are grounded in deliberate considerations. Sometimes they might learn from the other version, while other times they might just respect their differences as part of the diversity of human speech perception.

5.3.3 Suggestions for Improvement. To better handle the complexity of annotation, participants offered valuable suggestions. First, noting that audio segmentation could cause annotators to miss stuttering moments, Kerrigan suggested reminding annotators to listen beyond the timestamped segments and to examine the actual audio waveform.

Second, participants cautioned against listing hard rules or heuristics, because these rules often add complexity to understand stuttering as an "organic and complex" phenomenon. Breathing, for instance, sparked discussions: while Charan had leveraged it to identify stuttering events, Jia noted it can be both an avoidance technique and part of stuttering itself. She discussed the nuances with Wu (PWS co-author), as follows:

Jia: I would not pay too much attention to the breathing because that is almost like avoidance behavior, or people subconsciously do that to overcome those stuttering block. I feel this probably will add complexity to the already so complex work.

Wu: But, as you said, if people are trying to avoid stuttering by using the breathing techniques, doesn't that mean that they are actually stuttering? Even if they could kind of cover it, it's actually a stutter.

Jia: That's the one-million-dollar question, right? Because stuttering is so organic and complex. Sometimes trying not to stutter, is actually part of stuttering.

The conversation above reveals stuttering as a complex mix of involuntary disruptions and intentional bodily control, which blurs the boundary between stuttered and non-stuttered speech.

To preserve such multiplicities and allow room for interpretation, Wu proposed an evolving annotation model to replace the traditional majority voting method that assigns the most common label as the final label (e.g., [58]).

"We'll have one person code the whole thing, and then one person review their coding, and then one more person review their coding... We can always add more people that we think has more experience, or more knowledge in this, if we want to."

This idea gained support from both SLP participants, who further suggested strategies to better preserve embodied cues. For instance, Jia and Kerrigan both agreed that videos could be incorporated into the review process, as embodied cues such as lip movements are often a crucial aspect of stuttering and can aid interpretation [99].

5.4 Evolution of Guidelines

We integrated all participant feedback and updated the guidelines. The resulting guidelines are summarized in Section 7, and their evolution is summarized in Table 6. In particular, we refined the definition of prolongation (from sound to syllable) to prevent confusion and account for non-speech sounds that might otherwise be overlooked in text annotations. We also added reminders about common mislabeling patterns, clarified ambiguous cases, incorporated participant-shared heuristics, and included notes on software usage. Crucially, we added notes to encourage reflexive practice and ensure that heuristics are not treated as prescriptive.

Despite these efforts, the interpretation of stuttered speech proves to be inherently subjective and contextual, whether in navigating the messiness of speech flow, segmenting audio, or interpreting stuttering events. To respect the speaker's ultimate authority in defining their speech patterns, we decided to gather feedback from

PWS data contributors on annotations of their own speech and ground these evaluations in potential downstream applications.

6 Evaluation Sessions

We recruited four PWS participants who had contributed their speech data and reviewed annotated samples of their recordings with them. Our sessions were guided by two questions:

- How do PWS speakers perceive our labeling framework?
- How do PWS speakers perceive different versions of annotations?

The goal of this evaluation is not to provide a final assessment of performance but to explore PWS speakers' perceptions of our guidelines. Accordingly, we emphasized thematic saturation in recruitment and analysis.

6.1 Methods

Building on our earlier co-design sessions, we conducted one-hour Zoom interviews with four PWS speakers who contributed speech data to our inclusive speech AI initiative, Adedotun, Amina, Tatianna, and Benji. Their speech was annotated by at least two annotators familiar with stuttered speech, including annotations from annotators who did not participate in the co-design sessions. The annotators received the original audio recordings and TextGrid files containing auto-generated transcripts and timestamps produced by Zoom. Following the previous SLP session, the annotators used Praat to perform the annotation.

We received oral informed consent from the participants to conduct and record the interview for analysis. During the interviews, we asked participants about their perceptions of our annotation framework, the variations in annotations produced by different annotators, and their views on how such differences might affect downstream applications. We designed the following activities to guide these sessions. First, we presented examples of our annotated speech and asked the participants' general perception. We then presented different versions of annotations and asked participants about their perceptions and preferences. To facilitate reflection, we selected five examples of their annotated speech and invited participants to share their thoughts about the different versions.

To better contextualize our discussions, we asked our participants to consider the following five key scenarios and share their desired outcome and any potential issues with the annotations: (1) customizable auto-captioning in video conferencing, (2) voice commands for smart speakers, (3) speech-to-text features (e.g., for texting), (4) use in therapy contexts, (5) educational materials (e.g., training SLP students or promoting public awareness of stuttering). These scenarios were selected based on our PWS researcher's lived experiences with stuttering and our team's research expertise in this area. We also encouraged participants to suggest additional scenarios. To conclude the session, we brainstormed with participants about alternative ways to represent stuttered speech, inviting them to share any ideas they might have. We recorded all the video meeting sessions and transcribed the conversations for qualitative analysis. Our analysis identified themes consistent across participants regarding their perceptions of the labeling approach, annotation challenges, and label expectations.

6.2 Findings

Participants responded positively to our labeling framework. They noted that the labels capture a significant portion of stuttering (Adedotun, Tatianna), represent a big improvement over previous datasets (Benji), and emphasized the importance of verbatim transcription for reflecting a key aspect of their identity (Amina) and educating others about stuttering (Adedotun, Amina, Tatianna). However, their feedback also reflected limitations in relying on a fixed set of labels to capture their embodied experiences. This challenge fostered an appreciation for context-dependent evaluations of label accuracy.

6.2.1 Limitations in Capturing Embodied Richness. Participants brought complex embodied experiences that go far beyond what single labels can capture. In many instances, Adedotun and Amina identified more stuttering events in both annotations we presented, such as small interjections like “uh”, “um”. These short, light events could be easily missed by listeners but both participants were able to identify them due to their familiarity with their own stuttering patterns. Participants also suggested additional heuristics based on their embodied experiences, citing cues such as changes in voice volume, speech rate, and the number of attempts made to push through moments of disfluency.

In many places, participants offered a nuanced interpretation that was hard for a single label to capture. For example, Amina was hesitant to apply the block label to an instance where she did not experience the block in her usual way:

“I don't know if it's a block per se, because I tend to interpret a block as something that happens before I pronounce a word, when I cannot pronounce a word at all. But here I was kind of able to pronounce the first part and then the second part later. So I don't know how to label this.”

Similarly, Benji highlighted the value of phonetic annotations, noting that people who stutter might “*transform sounds*” during repetitions, making each repetition sound slightly different.

Participants also brought intimate understanding of their speech patterns that clarified the nuances missed in external observation. A common instance is that stuttering often blends tension and avoidance behaviors, making events easy to misclassify. For example, Amina explained that she often injects small interjections to get over blocks, which could blur the boundary between the two labels. Tatianna echoed similar points, attributing the differences in annotations to the gap between how stuttering sounds and how it is experienced internally:

“I don't fault annotation one, because it definitely sounded as if it was the repetitions. However, I know that's how I sound like when I have a block, as opposed to trying to bounce a sound.”

Furthermore, participants explained that they have certain stuttering patterns, which are important for differentiating labels. For example, Tatianna learned over years that her blocks are manifested mainly through repetitions and prolongations, while Benji described his stuttering as consisting primarily of repetitions. As stuttering manifests differently for each individual [107], participants interpreted and prioritized certain labels based on what

aspects of their speech they find most meaningful. For example, Adedotun highlighted the value of labeling silences:

“The only label that I can think of [that could be added] is silence. Sometimes with stuttering, you’re on a block, but there’s no sound being produced.”

Considering the dynamic nature of label meaning, our discussions then shifted to how speech labels should be interpreted in downstream applications.

6.2.2 Situated Label Interpretation and Accuracy Expectations. Participants expressed a desire to customize representations of their speech for different scenarios, including both verbatim and smoothed-over versions. Their expectations of accuracy shifted depending on conversational goals, social norms, relationships with interlocutors, and their own speech patterns. They shared many scenarios where they prioritized conveying content over representing stuttering events, e.g., when using voice commands with smart speakers, speech-to-text features, or talking to people who are familiar with their stuttering patterns such as family and friends.

Participants also affirmed the value of verbatim transcription, particularly when speaking to those unfamiliar with stuttering. In such scenarios, including stuttering labels in transcriptions was seen as important for conveying personal identity and promoting stuttering awareness. Tatianna, for instance, recalled a past job interview where her stuttering was mistaken for a lack of competence. She wished her stuttering had been accurately documented so she could have evidence to advocate for herself. Adedotun explained how he imagined a stuttering-aware captioning system could help the listener better follow stuttered speech:

“[I] think for the person who’s listening, it is helpful to know the time [of stuttering], so whenever the person is transitioning to the next word, they’re aware that the stutter speech is either no longer happening, or is continuing to happen.”

Participants expressed mixed expectations for therapeutic use, reflecting their differing orientations and goals. Some participants placed greater value on specific stuttering subtypes, as therapists need to “*learn from those different ways of stuttering and better take care of patients who stutter*” (Adedotun), and to see their “*whole self*” (Amina). Nevertheless, the utility of labels remains nuanced, as each individual may have their own perspective on what it means to speak ‘better’ and it varies depending on contexts [23]. For example, Adedotun and Benji placed less emphasis on specific stuttering events as they both prioritize self-acceptance in therapy. However, Adedotun acknowledged that for those who prioritize fluency, the specific labels might be important “*because the different ways of stuttering, you can use different techniques, so I think it would be important to have those nuances.*” These comments highlight how the different utilities of labels and the varied orientations toward stuttering described by participants complicate the goal of annotation. Given the diversity within the stuttering community, we recognized that accounting for individual orientations and speech patterns presents an ongoing challenge for annotation design.

6.2.3 Embracing Multiplicity with Care. Participants demonstrated an understanding that human subjectivity is an inherent aspect of

speech perception, provided the work is approached with care. Amina expressed a strong preference for having professionals handle the annotations:

“I would prefer they [the annotators] are at least familiar with the stuttered speech and know the different ways that stuttering occurs, not just a random person on the street.”

This strong preference for professionals highlights annotation as a process of encoding human knowledge and perception, rather than just a mechanical or objective task. Echoing this point, Tatianna noted that speech perception is deeply shaped by listeners’ personal language background, including potential differences between fluent and disfluent speakers, as well as between native and non-native English users.

Recognizing the inherent subjectivity in human perception, Adedotun valued the inconsistent labels, seeing them as a means to promote awareness and educate others about the diversity of human speech perception. Putting himself in the listener’s position, he proposed ways to express ambiguity in speech perception by using mixed notations such as “/b::p, /b:/p, and /b=/p.” These perspectives suggest that annotation is better understood as a form of collective sense-making, an ongoing effort to approach stuttering not as a fixed phenomenon, but as a multifaceted process requiring continuous reflection, interpretation, and care.

6.3 Continuous Refinements After Evaluations

While the PWS data contributors were generally satisfied with the annotations produced by our guidelines, the evaluation sessions highlighted a persistent gap between PWS’s internal experiences and the observable signals perceived externally. This observation suggests that the annotation of stuttered speech requires an open-ended approach, allowing space for uncertainty and iterative refinements. We thus call for continuous involvement of community members and professionals as data stewards, with the power to quality control and update the annotations as our experiences and knowledge of stuttering evolve. At the time of writing, Kerrigan and Benji have assumed the stewardship role of reviewing and refining the annotations for all speakers. We anticipate, and are committed to, continuously revising our guidelines under the guidance of community data stewards.

7 Annotation Guidelines and General Challenges

Here we summarize the annotation guidelines we developed. Due to space limitation, the full guidelines are presented in Appendix A. We also show the major changes made through co-design sessions as we iteratively developed the guidelines. While our guidelines were derived from the lived experiences of stuttering, they are not intended to reduce or replace the involvement of PWS in the annotation process, but to recognize and highlight the expertise of disabled people in creating and annotating disability datasets. By sharing a list of inherent challenges in annotating stuttered speech, we argue for the necessity of active and continuous involvement of PWS in this process.

| Changes Made | Session | Example |
|--|---------|---|
| Refined definitions | PWS | Changed “ <i>Elongated sound</i> ” to “ <i>Elongated syllable</i> ” for the prolongation label definition. |
| Added reminders regarding common mislabeling | PWS | Added “ <i>Be mindful of the potential difference between what you hear and the underlying stuttering events.</i> ” |
| Clarified ambiguous cases | PWS | Clarified “ <i>For repeated short one syllable words that are hard to tell between sound vs word repetitions, we will just label them as word repetitions (unless it is very clear sound repetitions of the first sound).</i> ” |
| Included heuristics participants shared | PWS | Added “ <i>Listen to non-verbal breathing sounds in front or end of the sentences.</i> ” |
| Added guidance on the annotation software use | SLP | Added “ <i>Listen beyond the timestamped segments and examine the audio waveform.</i> ” |
| Added clarifications for mixed stuttering events | SLP | Added a subsection titled “ <i>multiple stutter</i> ” and provided examples to illustrate that “ <i>one word can contain multiple stuttering events</i> ”. |
| Added reminders to encourage annotator reflections | SLP | Added “ <i>Note that they are not prescriptive, but rather meant to guide deeper thought.</i> ” in the section of heuristics to identify stuttering events. |

Table 6: Summary of guideline updates during co-design.

7.1 Annotation Guidelines Summary

7.1.1 Stuttering Annotation Guidelines. This section defines basic stuttering events to be annotated and introduces strategies to identify and distinguish them based on PWS expertise.

(1) Basic Stuttering Events

- Block (/b): a blocking pause before or within a word;
- Prolongation (/p): an elongated syllable;
- Sound repetition ([/s): repeated sound (excluding single-syllable words);
- Word/phrase repetition ([/r): repeated word or phrase, including single syllable word repetitions;
- Interjection ([/i): common or individualized filler words.

(2) Mixed Stuttering Events: the description of multiple stuttering events within one word (e.g. a sound is first repeated then prolonged) and how to annotate them, sometimes recursively.

(3) Identification Heuristics: tips to capture the nuances in different stuttering events. For example, leveraging breathing sounds to detect small blocks; distinguishing prolongation and block based on airflow. Most heuristics were derived from the bodily experience of stuttering. Non-exhaustive and non-prescriptive, these heuristics are intended as a reminder to pay attention to non-verbal signals rather than hard rules.

(4) Software Use: examining the audio waveform rather than relying on the software’s auto-segmentation, as sometimes stuttering events or cues can be cut out from auto segments.

7.1.2 General Speech Transcription Guidelines. This section describes transcription practices non-specific to stuttering.

- **Consistency:** the transcription needs to match the speech verbatim, containing all the filler words, repetitions, and connecting words;
- **Special Cases:** various instructions on how to transcribe accents, dialects, numbers, symbols, and acronym, following the general rule of transcribing what is heard;
- **Private and Sensitive Information:** mark and redact sensitive information.

Table 6 shares major changes resulting from the discussions with different participants, providing insights into how the guidelines evolved throughout this study.

7.2 General Challenges with Stuttered Speech Annotations

Despite the comprehensiveness of our annotation guidelines, our findings also uncover inherent challenges in annotating stuttered speech:

- The complex, organic, and often contested nature of stuttering. Producing faithful annotations thus requires embodied knowledge from PWS and, in some cases, insight from the speakers themselves.
- The ambiguity and subjectivity of human perception of speech flow. Experimental evidence shows that even subtle alterations to a stuffed toy presented to human subjects can shift their perception of vowel sounds [39].
- The socio-technical influences. Audio segmentation and the design of annotation tools can distort stuttered speech and omit key cues such as breathing or non-speech sounds.

We present a more detailed list in Appendix B and share more examples in our supplementary files.

8 Discussion

Our work offers the first attempt to bring disability-first dataset practices into AI dataset annotation. While prior efforts of disability-first datasets often focus on the data collection process [36, 47, 63, 68, 88, 98], our study represents an important step in centering disabled people and their expertise throughout the AI development pipeline. Our guidelines differ from SLP frameworks developed to measure fluency and stuttering severity [83, 109], as well as from earlier labeling frameworks designed primarily to detect stuttering events [58]. Instead, our guidelines seek to capture the lived knowledge of stuttering from PWS. Next, we reflect on the lessons learned and discuss implications for stuttered speech datasets and accessibility datasets in general.

8.1 Centering Disability Expertise in AI Dataset Annotation

Our work demonstrates the importance of incorporating PWS’s embodied knowledge in AI data annotation. Such knowledge represents what Hartblay called *disability expertise* – “the particular knowledge that disabled people develop and enact about unorthodox configurations of agency, cultural norms, and relationships between selves, bodies, and the designed world” [38]. In our study, this expertise is reflected in both the embodied experience of stuttering and the awareness of community interests; both are essential for capturing stuttering and addressing annotation challenges, such as the need for situated evaluation. While stuttering can be particularly internalized and subjective, the importance of disability expertise in disability data annotation is shared with many other disabilities, as the interpretation of data often relies on embodied knowledge unique to disabled bodies. For instance, hearing non-signers may overlook critical linguistic elements in sign language that deaf people are familiar with, such as facial expressions [28, 97]. Sighted people often do not understand what information blind people want in visual descriptions [31, 90]. Neurotypical people may have a narrowed sense of what communication can look like for neurodivergent people, missing multisensory channels [3, 79].

Missing disability expertise affects not only the accuracy of specific labels but also risks embedding ableism within the underlying AI training paradigm. For example, silent blocks are often mislabeled as non-speech segments and trimmed during ASR model training, leading to frequent interruptions of PWS speakers during speech blocks. Beyond stuttering, non-speech cues are central to disabled communicative practices [41], including those of autistic people [3], deaf people [55], and people with aphasia [34]. Ignoring the long-tail communication patterns and needs thus reflects a broader pattern of *epistemic injustice* in AI [2, 111]. Fricker used the concept to describe the harm done to people in their capacity as a knower, excluding their perspective due to bias, prejudice, or structural inequality [29]. Historically, such epistemic injustice has led to decades of pushback from deaf communities against sign language technologies that misrepresent sign languages [28, 84]. Involving disabled people in AI data annotation should therefore be a central step to mitigate such injustice.

Moreover, the intimacy and complexity of embodied knowledge pose deeper challenges for AI models to capture the full richness of disability. In the context of stuttering, Sheehan proposed the iceberg theory of stuttering [89], noting that “stuttering is like an iceberg, with only a small part above the waterline and a much bigger part below.” From speakers’ perspectives, stuttering involves more than observable behaviors; it also encompasses cognitive and affective reactions, including feelings, bodily sensations, and cognitive efforts that arise when anticipating moments of stuttering [22]. Reflecting similar ideas, the stuttering community advocates for more diverse evaluation measures, such as spontaneity rather than fluency [22]. The gap between what the current set of labels can capture and the richness of stuttering experiences reflects a long-standing question in HCI: the distinction between subjective, internal human experience and external observation [59]. Considering this gap, disabled people should have the authority to shape how their experiences

are represented, interpreted, and evaluated in AI systems, a principle captured by the idea of disability-first. Next, we reflect on our practice of the disability-first principle and draw lessons for accessibility datasets more broadly.

8.2 Toward Disability-First Annotations

A key lesson we learned is the value of involving stakeholders with dense and diverse expertise. Prior research suggested recruiting disabled people is a key challenge in creating disability-first datasets [78]. The diversity of disabled experiences also makes exhausting population experiences nearly impossible [93, 96]. In the context of stuttering, exposure to one’s own stuttered speech may even trigger intense emotional responses [105]. Faced with these challenges, we carefully recruited participants who possess a wide range of expertise including embodied stuttering experiences, experiences as the organizers and advocates of the stuttering community, experiences with stuttered speech annotations, and technical proficiency. Still, we acknowledge that achieving a disability-first approach must remain an ongoing commitment, and we will continue involving community members and professionals in later annotation and stewardship work.

A further challenge lies in how to ensure sustained disability leadership in a mixed-ability research and development team, a common setting in related datasets efforts [31, 88, 98]. Throughout our study, we prioritized the participation of PWS speakers in shaping our understanding of stuttering, with non-PWS members serving primarily in supporting roles. Our team ensured that PWS participants, including our team member, were given sufficient space and time to share their perspectives during team meetings or study sessions. When we encountered inconsistencies in annotations during the co-design sessions, PWS always voiced their perspectives first, while non-PWS focused on providing supporting evidence. Meanwhile, our non-PWS team members’ positionality as non-stutterers, experienced UX researchers, and long-time disability advocates allowed them to ask targeted questions that helped the PWS participants articulate their feelings and sensation into scientific knowledge about stuttering. These clarifying questions from non-PWS members were instrumental to uncover deeper insights from the lived experiences and to practice reflexivity. All analysis results were reviewed by the whole team and verified by the PWS senior author to ensure we fairly represented PWS perspectives. We hope our experience can guide mixed-ability research teams to develop disability-first team dynamics in other tasks and settings.

While centered on stuttering, our disability-first approach has potential to inform the development of accessibility datasets more broadly. Similar to stuttered speech datasets, labeling issues are common in accessibility datasets and can benefit from the involvement of PWD [12, 31]. For example, Garg et al. benchmarked vision-language models using a re-annotated subset of VizWiz, informed by a survey of blind participants [31]. The survey allowed them to incorporate information that blind people want but that sighted crowdworkers had not included [31]. Similarly, as explained in Section 8.1, many accessibility tasks require deeply personal embodied knowledge that disabled people possess, such as sign language translation [12, 95, 97], and disabled people’s non-normative communication practices [3, 34, 41]. Given the diversity of accessibility

tasks, a key question for future work is to further understand the role of disability expertise in dataset annotation across different tasks. For instance, in tasks measured against WCAG guidelines, annotators may be considered qualified as long as they are familiar with those guidelines. However, disabled people should be involved to evaluate if WCAG is sufficient as the measure of accessibility [80].

Future work should also investigate the challenges that arise in nuanced contexts when implementing the disability-first principle. In many tasks, disabled people might have limited access to sensory information that is important to the task [33, 75], or their desired information might come into tension with other values such as privacy or making racial inference from photos [10, 37, 94]. In these situations, collaboration among people of mixed abilities and professionals is necessary. Crucially, while essential, the inclusion of disabled individuals is not a panacea [25, 91], especially as the notion of ‘meaningful access’ remains contested and evolving. For example, disabled people may lower their expectations due to the pervasive inaccessibility they have experienced in everyday life [25, 94]. Disability-first is therefore more than just seeking user feedback; it involves equipping the communities with the resources, power, and support they need – via training, advocacy, and sustained investment – so that they can take ownership of the technologies that affect them.

8.3 Embracing Multiplicity Throughout the AI Development Pipeline

Ultimately, we advocate for acknowledging and working with the ambiguity and context dependencies of human data across the AI pipeline. Our findings show that some annotation challenges are inherent to human perception and cannot be fully resolved with a single guideline. These challenges echo recent work that treats data practices as reflexive practices [16] and frames inconsistencies as sites for reflection [65]. The case of stuttering provides yet another example showing that label inconsistencies are not always problems to be ‘settled’; rather, they reflect the diversity of human perception and can prompt ongoing, reflexive, and collaborative efforts to better understand human experiences. Similar to stuttering, many traditional accessibility tasks inherently involve ambiguity and subjective interpretation, such as sign language translation [12] and emotion recognition [8]. Subjectivity even appears in traditionally more ‘objective’ tasks such as visual question answering (VQA) [94, 110]. Although VQA has traditionally been treated as a task with a single ground-truth label [4], recent work highlights the inherent subjectivity in visual information seeking, showing that blind users integrate multiple perspectives to shape their visual understanding, whether with AI tools or humans [94].

Although task-specific research is needed, a multiplicity-aware perspective offers a valuable paradigm shift in data practices related to human perception. It calls researchers to be sensitive to practices erasing the diversity of human perception, such as rigid or organizational control [1] and majority voting [24]. Recent work has highlighted reflexive practices drawn from interpretivist traditions, such as encouraging data workers to reflect on their positionality [16] and drawing on qualitative research methods to aid in data interpretation [103]. Reflecting similar ideas, our SLP participants

proposed a stewardship model in which a more experienced member oversees the annotation process. Building on this emphasis on interpretive labeling, tools could further support annotators in understanding differing perspectives, such as systems that predict the diversity of human responses [108]. Such systems could also benefit users by revealing the different angles that informed labeling.

Specific to stuttering applications, a key lesson we learned is the situated meaning of speech labels. While PWS participants showed strong intuition on which label ‘best’ captures the speech patterns, they were also conscious about the space of interpretation and the fairness risk when deterministically assigning one label as the “ground truth” [69]. The need to evaluate labels within the context of downstream applications has been highlighted by our participants. For example, annotations might prioritize content accuracy in certain scenarios such as voice commands and speech-to-texting, while focusing on accurate stuttering event labels in therapy or educational contexts. Additionally, PWS might prioritize the accuracy of certain labels according to their speech patterns. These varying priorities suggest that annotation should reflect the goals of the intended application rather than enforcing uniformity.

Another takeaway from our study is that interpretation of speech labels should extend beyond the moment of annotation. As PWS participants imagined the use of their annotated speech, the interpretation of labels is never final but continuously evolving; for example, inaccurate annotations of stuttering type might still be useful in helping conversation partners become aware of the existence of stuttering, which could enhance listeners’ comprehension of stuttered speech [15]. This way, the meaning of labels is co-constructed through their evolving use, rather than being fixed within datasets. Designing for such ongoing interpretations of labels requires approaches that encourage open, multiple interpretations [87]. For example, systems could leave interpretations open to users using ambiguous labels such as /b::/p as Adedotun suggested. We can even envision a broader ecosystem that supports the ongoing collection of annotations from PWS. For instance, a self-annotation application could help PWS with self-therapy or communicate with SLPs, while the annotations could be shared to improve speech models or used for fine-tuning with informed consent. We encourage future work to explore continuous, participatory approaches that sustain annotation efforts over time and fully integrate diverse stuttering experiences into AI development pipelines.

8.4 Limitations and Future Work

This study has several limitations. First, our study relies on a relatively small number of participants. Despite our efforts in engaging multiple perspectives and community voices, our perspectives may be skewed toward people with higher socio-economic backgrounds and greater technical literacy, and under-represent those with multiple disabilities [102]. Second, although the stuttering community has been exploring new ways to understand stuttering, our perspectives, including those of PWS, are inevitably influenced by pre-existing frameworks, such as those from SLPs. Part of our guidelines was still built off existing labeling frameworks. Given these two aspects, we argue that the guidelines should continue to evolve alongside people’s experiences and perspectives. Future work should involve more community members to build upon and

refine our efforts. Third, the speech data we collected and used did not cover variation across English dialects and other languages. Additional challenges could surface when extending our approach to more diverse speech patterns.

9 Conclusion

We present a case study exploring a disability-first approach to annotating AI datasets in the context of stuttered speech. Through a collaborative effort involving HCI researchers, PWS AI professionals, SLP professionals, and PWS data contributors, we explored the challenges of annotating stuttered speech and co-designed annotation guidelines grounded in the lived experiences of PWS. Our findings show the critical role of embodied knowledge in shaping more responsible annotation practices. We aim for this work to inspire broader efforts to center disabled knowledge and experiences throughout the AI development pipeline.

Acknowledgments

We thank all participants – Charan Sridhar, Rong Gong, Jia Bin, Julia Kerrigan, Adedotun Bello, Amina Kobenova, Tatianna Vassilopoulos, and Benji Schussheim – for their time and valuable insights. We are particularly grateful to our participants for consenting to the use of their real identities in this publication – an act that reflects remarkable trust and confidence in our work. We also thank the Authentic User Experience Lab at the University of California, Santa Cruz, especially Professor Norman Makoto Su, for stimulating discussions and collaboration on the future development of this work. Finally, we extend our heartfelt gratitude to all speech data contributors for entrusting us with their data that grounded this work. This work was supported by NSF Award #2427710, the Patrick J. McGovern Foundation, and Borealis Philanthropy's Disability Inclusion Fund.

References

- [1] Nuredin Ali Abdelkadir, Tianling Yang, Shivani Kapania, Meron Estefanos, Fasca Berhane Gebrekidan, Zecharias Zelalem, Messai Ali, Rishan Berhe, Dylan Baker, Zeerak Talat, Milagros Miceli, Alex Hanna, and Timnit Gebru. 2025. The Role of Expertise in Effectively Moderating Harmful Social Media Content. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 227, 21 pages. doi:10.1145/3706598.3714010
- [2] Leah Hope Ajmani, Jasmine C. Foriest, Jordan Taylor, Kyle Pittman, Sarah Gilbert, and Michael Ann DeVito. 2024. Whose Knowledge is Valued? Epistemic Injustice in CSCW Applications. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 523 (Nov. 2024), 28 pages. doi:10.1145/3687062
- [3] Meryl Alper. 2018. Inclusive sensory ethnography: Studying new media and neurodiversity in everyday life. *New Media & Society* 20, 10 (2018), 3560–3579. <https://doi.org/10.1177/1461444818755394>
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [5] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 167–176. doi:10.1145/3351095.3372859
- [6] Ashita Batra, Mannas Narang, Neeraj Kumar Sharma, and Pradip K. Das. 2025. Boli: A dataset for understanding stuttering experience and analyzing stuttered speech. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–4. doi:10.1109/ICASSP49660.2025.10888349
- [7] Sebastian Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Noeth, and Korbinian Riedhammer. 2022. KSoF: The Kassel State of Fluency Dataset – A Therapy Centered Dataset of Stuttering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric B  chet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H  l  ne Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1780–1787. <https://aclanthology.org/2022.lrec-1.189/>
- [8] Andrew Begel, John Tang, Sean Andrist, Michael Barnett, Tony Carbary, Piali Choudhury, Edward Cutrell, Alberto Fung, Sasa Junuzovic, Daniel McDuff, Kael Rowan, Shibashankar Sahoo, Jennifer Frances Waldern, Jessica Wolk, Hui Zheng, and Annuska Zolyomi. 2020. Lessons Learned in Designing AI for Autistic Adults. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (Virtual Event, Greece) (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 6 pages. doi:10.1145/3373625.3418305
- [9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FACeT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922
- [10] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 375, 19 pages. doi:10.1145/3411764.3445498
- [11] Paul Boersma and Vincent Van Heuven. 2001. Speak and unSpeak with PRAAT. *Glott International* 5, 9/10 (2001), 341–347.
- [12] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective. *ACM Trans. Access. Comput.* 14, 2, Article 7 (July 2021), 45 pages. doi:10.1145/3436996
- [13] Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring Collection of Sign Language Videos through Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 514 (Nov. 2022), 24 pages. doi:10.1145/3555627
- [14] Virginia Braun and Victoria Clarke. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research* 21, 1 (2021), 37–47. <https://doi.org/10.1002/capr.12360>
- [15] C. T. Byrd, M. McGill, Z. Gkalitsiou, and C. Cappellini. 2017. The effects of self-disclosure on male and female perceptions of individuals who stutter. *American Journal of Speech-Language Pathology* 26, 1 (2017), 69–80. https://doi.org/10.1044/2016_AJSLP-15-0164
- [16] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 572, 19 pages. doi:10.1145/3491102.3501998
- [17] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and inter-coder reliability and agreement. *Sociological methods & research* 42, 3 (2013), 294–320. <https://doi.org/10.1177/0049124113500475>
- [18] Yang Trista Cao, Kyle Seelman, Kyungjun Lee, and Hal Daum   III. 2022. What's Different between Visual Question Answering for Machine "Understanding" Versus for Accessibility?. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 1025–1034. doi:10.18653/v1/2022.aacl-main.75
- [19] Christopher Constantino, Patrick Campbell, and Sam Simpson. 2022. Stuttering and the social model. *Journal of Communication Disorders* 96 (2022), 106200. <https://doi.org/10.1016/j.jcomdis.2022.106200>
- [20] Christopher D Constantino. 2018. What can stutterers learn from the neurodiversity movement?. In *Seminars in Speech and Language*, Vol. 39. Thieme Medical Publishers, 382–396. doi:10.1055/s-0038-1667166
- [21] Christopher Dominick Constantino. 2023. Fostering positive stuttering identities using stutter-affirming therapy. *Language, Speech, and Hearing Services in Schools* 54, 1 (2023), 42–62. https://doi.org/10.1044/2022_LSHSS-22-00038
- [22] Christopher Dominick Constantino, Naomi Eichorn, Eugene H Buder, J Gayle Beck, and Walter H Manning. 2020. The speaker's experience of stuttering: Measuring spontaneity. *Journal of Speech, Language, and Hearing Research* 63, 4 (2020), 983–1001. https://doi.org/10.1044/2019_JSLHR-19-00068
- [23] Christopher Dominick Constantino, Walter H Manning, and Susan Naomi Nordstrom. 2017. Rethinking covert stuttering. *Journal of Fluency Disorders* 53 (2017), 26–40. <https://doi.org/10.1016/j.jfludis.2017.06.001>
- [24] Aida Mostafazadeh Davani, Mark D  az, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10

- (2022), 92–110. https://doi.org/10.1162/tacl_a_00449
- [25] Maartje De Meulder. 2021. Is “good enough” good enough? Ethical and responsible development of sign language technologies. In *Proceedings of the 1st international workshop on automatic translation for signed and spoken languages (AT4SSL)*. 12–22. <https://aclanthology.org/2021.mtsummit-at4ssl/2/>
- [26] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems* 36 (2023), 76893–76907. https://proceedings.neurips.cc/paper_files/paper/2023/file/f29c8f8b4996a4a453ef366cf496354-Paper-Datasets_and_Benchmarks.pdf
- [27] Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. 2024. Systemic Biases in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*. 54–65. <http://www.lrec-conf.org/proceedings/lrec-coling-2024/signlang/pdf/2024.signlang-1.6.pdf>
- [28] Michael Erard. 2017. Why Sign-Language Gloves Don’t Help Deaf People. <https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/>. Retrieved June 3, 2025.
- [29] Miranda Fricker. 2017. Evolving concepts of epistemic injustice. In *The Routledge handbook of epistemic injustice*. Routledge, 53–60.
- [30] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. “I wouldn’t say offensive but...”: Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT ’23)*. Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989
- [31] Kapil Garg, Xinru Tang, Jimin Heo, Dwayne R Morgan, Darren Gergle, Erik B Sudderth, and Anne Marie Piper. 2025. “It’s trained by non-disabled people”: Evaluating How Image Quality Affects Product Captioning with VLMs. *arXiv preprint arXiv:2511.08917* (2025).
- [32] Rong Gong, Hongfei Xue, Lezhi Wang, Xin Xu, Qisheng Li, Lei Xie, Hui Bu, Shaomei Wu, Jiaming Zhou, Yong Qin, et al. 2024. AS-70: A mandarin stuttered speech dataset for automatic speech recognition and stuttering event detection. *arXiv preprint arXiv:2406.07256* (2024).
- [33] Steven M Goodman, Ping Liu, Dhruv Jain, Emma J McDonnell, Jon E Froehlich, and Leah Findlater. 2021. Toward user-driven sound recognizer personalization with people who are d/deaf or hard of hearing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–23. doi:10.1145/3463501
- [34] Charles Goodwin. 2004. A competent speaker who can’t speak: The social life of aphasia. *Journal of Linguistic Anthropology* 14, 2 (2004), 151–170. <https://doi.org/10.1525/jlin.2004.14.2.151>
- [35] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (Denver, Colorado, USA) (CHI ’17)*. Association for Computing Machinery, New York, NY, USA, 3511–3522. doi:10.1145/3025453.3025781
- [36] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizviz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617. http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html
- [37] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 543–554. doi:10.1145/3461702.3462620
- [38] Cassandra Hartblay. 2020. Disability expertise: claiming disability anthropology. *Current Anthropology* 61, S21 (2020), S26–S36. doi:10.1086/705781
- [39] Jennifer Hay and Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48, 4 (2010). doi:10.1515/LING.2010.027
- [40] Peter A Heeman, Andy McMillin, and J Scott Yaruss. 2006. An annotation scheme for complex disfluencies. In *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006-ICSLP*. International Speech Communication Association, 1081–1084. doi:10.21437/Interspeech.2006-55
- [41] Jon Henner and Octavian Robinson. 2023. Unsettling languages, unruly body-minds: A crip linguistics manifesto. *Journal of Critical Study of Communication and Disability* 1, 1 (2023), 7–37. <https://doi.org/10.48516/jcsd.2023vol1iss1.4>
- [42] Peter Howell, Stephen Davis, and Jon Bartrip. 2009. The University College London Archive of Stuttered Speech (UCLASS). *Journal of speech, language, and hearing research* 52, 2 (2009), 556–569. [https://doi.org/10.1044/1092-4388\(2009/07-0129\)](https://doi.org/10.1044/1092-4388(2009/07-0129))
- [43] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5491–5501. doi:10.18653/v1/2020.acl-main.487
- [44] Jackson, Eric and Quesal, Robert and Yaruss, J. Scott. 2012. What Is Stuttering: Revisited. <https://ahn.mnsu.edu/services-and-centers/center-for-communication-sciences-and-disorders/services/stuttering/professional-education/convention-materials/archive-of-online-conferences/isad2012/what-is-stuttering-revisited/>. Retrieved September, 2025.
- [45] James Madison University. 2025. Transcription Guidelines. <https://www.jmu.edu/ods/calm/caption-on/diy-captioning/transcription-guidelines.shtml>. Accessed September 2025.
- [46] Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka Jha, and Riya Chanduka. 2020. IncluSet: A Data Surfacing Repository for Accessibility Datasets (ASSETS ’20). Association for Computing Machinery, New York, NY, USA, Article 72, 4 pages. doi:10.1145/3373625.3418026
- [47] Rie Kamikubo, Utkarsh Dwivedi, and Hernisa Kacorri. 2021. Sharing Practices for Datasets Related to Accessibility and Aging. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (Virtual Event, USA) (ASSETS ’21)*. Association for Computing Machinery, New York, NY, USA, Article 28, 16 pages. doi:10.1145/3441852.3471208
- [48] Rie Kamikubo, Farnaz Zamiri Zeraati, Kyungjun Lee, and Hernisa Kacorri. 2024. AccessShare: Co-designing Data Access and Sharing with Blind People. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (St. John’s, NL, Canada) (ASSETS ’24)*. Association for Computing Machinery, New York, NY, USA, Article 60, 16 pages. doi:10.1145/3663548.3675612
- [49] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI ’23)*. Association for Computing Machinery, New York, NY, USA, Article 133, 15 pages. doi:10.1145/3544548.3580645
- [50] Rhea Kapur and Elisa Kreiss. 2024. Reference-based metrics are biased against blind and low-vision users’ image description preferences. In *Proceedings of the Third Workshop on NLP for Positive Impact*. 308–314. doi:10.18653/v1/2024.nlp4pi-1.26
- [51] Gunay Kazimzade and Milagros Miceli. 2020. Biased Priorities, Biased Outcomes: Three Recommendations for Ethics-oriented Data Annotation Practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (New York, NY, USA) (AI/ES ’20)*. Association for Computing Machinery, New York, NY, USA, 71. doi:10.1145/3375627.3375809
- [52] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAccT ’24)*. Association for Computing Machinery, New York, NY, USA, 1672–1681. doi:10.1145/3630106.3658996
- [53] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2021. Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2986–2999. doi:10.1109/TASLP.2021.3110146
- [54] Annelies Kusters, Maartje De Meulder, Dai O’Brien, et al. 2017. Innovations in deaf studies: Critically mapping the field. *Innovations in deaf studies: The role of deaf scholars* 12 (2017), 1–53.
- [55] Annelies Kusters, Massimiliano Spotti, Ruth Swanwick, and Elina Tapio. 2017. Beyond languages, beyond modalities: Transforming the study of semiotic repertoires. *International Journal of Multilingualism* 14, 3 (2017), 219–232. <https://doi.org/10.1080/14790718.2017.1321651>
- [56] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda. 2002. Lightly supervised and unsupervised acoustic model training. *Comput. Speech Lang.* 16, 1 (Jan. 2002), 115–129. doi:10.1006/csla.2001.0186
- [57] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16. doi:10.1145/3544548.3581224
- [58] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P. Bigham. 2021. SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6798–6802. doi:10.1109/ICASSP39728.2021.9413520
- [59] Lucian Leahu, Steve Schwenk, and Phoebe Sengers. 2008. Subjective objectivity: negotiating emotional meaning. In *Proceedings of the 7th ACM conference on Designing interactive systems*. 425–434. doi:10.1145/1394445.1394491
- [60] Jingjin Li, Qisheng Li, Rong Gong, Lezhi Wang, and Shaomei Wu. 2025. Our Collective Voices: The Social and Technical Values of a Grassroots Chinese Stuttered Speech Dataset (FAccT ’25). Association for Computing Machinery, New York, NY, USA, 2768–2783. doi:10.1145/3715275.3732179
- [61] Jingjin Li, Peiyao Liu, Rebecca Lietz, Ningjing Tang, Norman Makoto Su, and Shaomei Wu. 2025. Govern With, Not For: Understanding the Stuttering Community’s Preferences and Goals for Speech AI Data Governance in the US and China. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*,

- Vol. 8. 1548–1560. <https://doi.org/10.1609/aies.v8i2.36654>
- [62] Jingjin Li, Shaomei Wu, and Gilly Leshed. 2024. Re-envisioning Remote Meetings: Co-designing Inclusive and Empowering Videoconferencing with People Who Stutter. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) (*DIS '24*). Association for Computing Machinery, New York, NY, USA, 1926–1941. doi:10.1145/3643834.3661533
- [63] Qisheng Li and Shaomei Wu. 2024. “I Want to Publicize My Stutter”: Community-led Collection and Curation of Chinese Stuttered Speech Data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–27. doi:10.1145/3687014
- [64] Qisheng Li and Shaomei Wu. 2024. Towards Fair and Inclusive Speech Recognition for Stuttering: Community-led Chinese Stuttered Speech Dataset Creation and Benchmarking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9. doi:10.1145/3613905.3650950
- [65] Cindy Kaiying Lin and Steven J Jackson. 2023. From bias to repair: Error as a site of collaboration and negotiation in applied data science work. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32. doi:10.1145/3579607
- [66] Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, Philip C Nelson, et al. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech*, Vol. 2021. 4833–4837. doi:10.21437/Interspeech.2021-697
- [67] Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, and Cecily Morrison. 2024. Explaining CLIP’s Performance Disparities on Data from Blind/Low Vision Users. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12172–12182.
- [68] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. 2021. Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10818–10828.
- [69] Anna P. Meyer, Yea-Seul Kim, Loris D’Antoni, and Aws Albarghouthi. 2025. Perceptions of the Fairness Impacts of Multiplicity in Machine Learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (*CHI '25*). Association for Computing Machinery, New York, NY, USA, Article 524, 15 pages. doi:10.1145/3706598.3713524
- [70] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 34 (Jan. 2022), 14 pages. doi:10.1145/3492853
- [71] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 115 (Oct. 2020), 25 pages. doi:10.1145/3415186
- [72] Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. Documenting Data Production Processes: A Participatory Approach for Data Work. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 510 (Nov. 2022), 34 pages. doi:10.1145/3555623
- [73] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (*FAccT '21*). Association for Computing Machinery, New York, NY, USA, 161–172. doi:10.1145/3442188.3445880
- [74] Kowshik Siva Sai Motepalli, Vamshiraghushimha Narasinga, Harsha Pathuri, Hina Khan, Sangeetha Mahesh, Ajish K. Abraham, and Anil Kumar Vuppala. 2023. Stuttering Detection Application. In *Interspeech 2023*. 680–681.
- [75] Annika Muehlbradt and Shaun K. Kane. 2022. What’s in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Trans. Access. Comput.* 15, 1, Article 6 (March 2022), 32 pages. doi:10.1145/3507659
- [76] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 94, 16 pages. doi:10.1145/3411764.3445402
- [77] NIDCD Information Clearinghouse. [n.d.]. Stuttering. <https://www.nidcd.nih.gov/health/stuttering>. Retrieved June, 2025.
- [78] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (*FAccT '21*). Association for Computing Machinery, New York, NY, USA, 52–63. doi:10.1145/3442188.3445870
- [79] Amit Pinchevski and John Durham Peters. 2016. Autism and new media: Disability between technology and society. *New Media & Society* 18, 11 (2016), 2507–2523. <https://doi.org/10.1177/1461444815594441>
- [80] Christopher Power, André Freire, Helen Petrie, and David Swallow. 2012. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) (*CHI '12*). Association for Computing Machinery, New York, NY, USA, 433–442. doi:10.1145/2207676.2207736
- [81] Kerri Prinos, Neal Patwari, and Cathleen A. Power. 2024. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) (*FAccT '24*). Association for Computing Machinery, New York, NY, USA, 1245–1254. doi:10.1145/3630106.3658969
- [82] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 28492–28518. <https://proceedings.mlr.press/v202/radford23a.html>
- [83] Nan Bernstein Ratner and Brian MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders* 56 (2018), 69–80. <https://doi.org/10.1016/j.jfludis.2018.03.002>
- [84] r/deaf. [n.d.]. NEW total ban on research affective immediately! https://www.reddit.com/r/deaf/comments/1i4gk9n/new_total_ban_on_research_affective_immediately/ Retrieved June 5, 2025.
- [85] Morgan Klaus Scheuerman, Kandra Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We’ve Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 58 (May 2020), 35 pages. doi:10.1145/3392866
- [86] Morgan Klaus Scheuerman, Allison Woodruff, and Jed R. Brubaker. 2025. How Data Workers Shape Datasets: The Role of Positionality in Data Collection and Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW300 (Oct. 2025), 42 pages. doi:10.1145/3757481
- [87] Phoebe Sengers and Bill Gaver. 2006. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) (*DIS '06*). Association for Computing Machinery, New York, NY, USA, 99–108. doi:10.1145/1142405.1142422
- [88] Tanusree Sharma, Abigale Stangl, Lotus Zhang, Yu-Yun Tseng, Inan Xu, Leah Findlater, Danna Gurari, and Yang Wang. 2023. Disability-First Design and Creation of A Dataset Showing Private Visual Information Collected With People Who Are Blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 51, 15 pages. doi:10.1145/3544548.3580922
- [89] Joseph G Sheehan. 1970. Stuttering: Research and therapy. (1970).
- [90] Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. 2020. “I Hope This Is Helpful”: Understanding Crowdworkers’ Challenges and Motivations for an Image Description Task. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 105 (Oct. 2020), 26 pages. doi:10.1145/3415176
- [91] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) (*EAAMO '22*). Association for Computing Machinery, New York, NY, USA, Article 1, 6 pages. doi:10.1145/3551624.3555285
- [92] Charan Sridhar and Shaomei Wu. 2025. J-j-j-just Stutter: Benchmarking Whisper’s Performance Disparities on Different Stuttering Patterns. In *Interspeech 2025*. 3753–3757. doi:10.21437/Interspeech.2025-2700
- [93] Xinru Tang. 2024. Why is Accessibility So Hard? Insights From the History of Privacy. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) (*CSCW Companion '24*). Association for Computing Machinery, New York, NY, USA, 362–368. doi:10.1145/3678884.3681876
- [94] Xinru Tang, Ali Abdolrahmani, Darren Gergle, and Anne Marie Piper. 2025. Everyday Uncertainty: How Blind People Use GenAI Tools for Information Access. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (*CHI '25*). Association for Computing Machinery, New York, NY, USA, Article 63, 17 pages. doi:10.1145/3706598.3713433
- [95] Xinru Tang, Xiang Chang, Nuoran Chen, Yingjie (MaoMao) Ni, RAY LC, and Xin Tong. 2023. Community-Driven Information Accessibility: Online Sign Language Content Creation within d/Deaf Communities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 50, 24 pages. doi:10.1145/3544548.3581286
- [96] Xinru Tang, Gabriel Lima, Jiang Jiang, Li, Lucy Simko, and Yixin Zou. 2025. Beyond “Vulnerable Populations”: A Unified Understanding of Vulnerability From A Socio-Ecological Perspective. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW037 (May 2025), 30 pages. doi:10.1145/3710935
- [97] Xinru Tang and Anne Marie Piper. 2026. Reimagining Sign Language Technologies: Analyzing Translation Work of Chinese Deaf Online Content Creators. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*

- (Barcelona, Spain) (CHI '26). Association for Computing Machinery, New York, NY, USA, 20 pages. doi:10.1145/3772318.3790624
- [98] 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. doi:10.1145/3441852.3471225
- [99] Seth E Tichenor and J Scott Yaruss. 2019. Stuttering as defined by adults who stutter. *Journal of Speech, Language, and Hearing Research* 62, 12 (2019), 4356–4369. https://doi.org/10.1044/2019_JSLHR-19-00137
- [100] Seth E Tichenor and J Scott Yaruss. 2021. Variability of stuttering: Behavior and impact. *American Journal of Speech-Language Pathology* 30, 1 (2021), 75–88. https://doi.org/10.1044/2020_AJSLP-20-00112
- [101] Mindy Tran, Xinru Tang, Adryana Hutchinson, Adam J Aviv, and Yixin Zou. 2026. Toward Inclusive Security and Privacy for Deaf and Hard-of-Hearing People: A Community-Based Interview Study. In *2026 IEEE Symposium on Security and Privacy (S&P)*. IEEE.
- [102] Ana Rita Valente, Rufael Marew, Hawau Olamide Toyin, Hamdan Al-Ali, Anelise Bohnen, Inma Becerra, Elsa Marta Soares, Goncalo Leal, and Hanan Aldarmaki. 2025. Clinical Annotations for Automatic Stuttering Severity Assessment. *arXiv preprint arXiv:2506.00644* (2025).
- [103] Ruyuan Wan, Haonan Wang, Ting-Hao Kenneth Huang, and Jie Gao. 2025. From Noise to Nuance: Enriching Subjective Data Annotation through Qualitative Analysis. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Siyan Li, Michael Madaio, Jack Wang, Sherry Tongshuang Wu, Ziang Xiao, and Diyi Yang (Eds.). Association for Computational Linguistics, Suzhou, China, 240–254. doi:10.18653/v1/2025.hcinp-1.20
- [104] Meredith Whittaker, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Elizabeth Kazunas, Mara Mills, Meredith Ringel Morris, Joy Lisi Rankin, Emily Rogers, Marcel Salas, and Sarah Myers West. 2019. Disability, Bias & AI Report. *AI Now Institute* (20 11 2019).
- [105] Shaomei Wu. 2023. "The World is Designed for Fluent People": Benefits and Challenges of Videoconferencing Technologies for People Who Stutter. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 782, 17 pages. doi:10.1145/3544548.3580788
- [106] Shaomei Wu, Jingjin Li, and Gilly Leshed. 2024. Finding My Voice over Zoom: An Autoethnography of Videoconferencing Experience for a Person Who Stutters. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 916, 16 pages. doi:10.1145/3613904.3642746
- [107] Ehud Yairi. 2007. Subtyping stuttering I: A review. *Journal of fluency disorders* 32, 3 (2007), 165–196. <https://doi.org/10.1016/j.jfludis.2007.04.001>
- [108] Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual question answer diversity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 184–192. <https://doi.org/10.1609/hcomp.v6i1.13341>
- [109] J Scott Yaruss. 1997. Clinical measurement of stuttering behaviors. *Contemporary Issues in Communication Science and Disorders* 24, Spring (1997), 27–38. https://doi.org/10.1044/cicsd_24_S_27
- [110] Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. 2025. Semantic and Expressive Variations in Image Captions Across Languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29667–29679.
- [111] Anon Ymous, Katta Spiel, Os Keyes, Rua M. Williams, Judith Good, Eva Hornecker, and Cynthia L. Bennett. 2020. "I am just terrified of my future" — Epistemic Violence in Disability Related Technology Research. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3334480.3381828
- [112] Ben Zefeng Zhang, Tianling Yang, Oliver Haimson, and Michaelanne Thomas. 2024. Conceptualizing Precision Labor in Artificial Intelligence Training. *AoIR Selected Papers of Internet Research* (2024). <https://doi.org/10.5210/spir.v2024i0.14079>
- [113] Ben Zefeng Zhang, Tianling Yang, Milagros Miceli, Oliver L. Haimson, and Michaelanne Thomas. 2025. The Making of Performative Accuracy in AI Training: Precision Labor and Its Consequences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 1203, 19 pages. doi:10.1145/3706598.3713112
- [114] Xin Zhang, Iván Vallés-Pérez, Andreas Stolcke, Chengzhu Yu, Jasha Droppo, Olabanji Shonibare, Roberto Barra-Chicote, and Venkatesh Ravichandran. 2022. Stutter-tts: Controlled synthesis and improved recognition of stuttered speech. *arXiv preprint arXiv:2211.09731* (2022).
- [115] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173789

A Full Annotation Guidelines

In general, the speech need to be transcribed verbatim, meaning that, all repetitions, interjections (e.g. uh, wow, oh), filler words (e.g. you know, like), and even stuttered sounds need to be transcribed in the text as they are. The following sections will break down the specific requirements and standards for transcription.

1. Stuttering Annotation

1.1 Basic annotation

For stuttering or other non-stuttering disfluencies such as self-correction or repeated words, transcribe what you heard verbatim. When the disfluency is induced by stuttering, also mark the type of stutter using the following annotations described in Table 7.

Notes:

1.2 Multiple stutter

One word can contain multiple stuttering events, thus be annotated with more than one stuttering code. A few examples:

- “[m-m-m-]/sm/py”: the “m” sound in “my” is repeated three times and prolonged;
- “I [uh uh uh]/r/i /bwork”: “uh” as a stuttering interjection was repeated three times, and followed by a block before the speaker said “work”.

1.3 Single syllable word repetition

For repeated short one syllable words that are hard to tell between sound vs word repetitions, we will just label them as word repetitions (unless it is very clear sound repetitions of the first sound).

1.4 Be mindful of the potential difference between what you hear and the underlying stuttering events.

Some heuristics our team observed in our pilot stage are listed below. Note that they are not prescriptive, but rather meant to guide deeper thought.

- (1) Listen to non-verbal breathing sounds in front or end of the sentences. This is particularly important for detecting small blocks.
- (2) Distinguish involuntary repetition and block then backtracking. Sound repetition is involuntary but the repetition from block and backtracking is intentional. We should label it as a block.
- (3) Distinguish prolongation vs block and trying to make the sound during block. The airflow is continuous in prolongation, but choppy in blocks. If the air is not flowing, we should label it as a block.

1.5. listen beyond the timestamped segments and examine the audio waveform. Sometimes stuttering events or cues could be missed outside of the segments.

2. Consistency

The transcript needs to be consistent with the speech, even if the speaker makes grammatical errors. Keep all the connecting words such as “and”, “but”, and “so”.

3. Accents and Dialects

Do not transcribe accents or dialects by pronunciation unless there is a common spelling for the accented word (e.g. “y’all”). Use contractions only if they are used by the speaker. If the speaker mispronounced a word, use the intended word in the transcription, except when the speaker mis-spoke the entire word.

For example, when the speaker intended to say “royal palace” but said it as “royal place”, you should transcribe it verbatim as “royal place”. But if the speaker said “royal plaze”, you should transcribe it as “royal place”.

4. Sensitive Information

Mark sensitive information, such as names, age, specific occupation, specific places, in < >. For example: I am <Katie> and I live at <Athens, Ohio>.

5. Numbers and Symbols

Transcribe numbers in English rather than Arabic numbers. E.g. “We will meet at two thirty”.

Symbols should be transcribed by the sound, too. E.g., when the speakers mention an email address or website, it should be transcribed as “info at university dot org”, not “info@university.org”.

6. Acronyms

Transcribe acronyms as it was said, for example, NASA will be transcribed as “NASA”, not “N A S A”.

If the acronym is spelled out letter by letter to emphasize the spelling, add a hyphen in between letters. For example, “I am part of the A-C-T, ACT program.”

If it is a common acronym that is not spelled out by the speaker, write it without hyphen. E.g. “I have ADHD”, or “AI is changing the world”.

B Challenges in Annotations

This list is not prescriptive or intended to be exhaustive. The main goal to list these notes and thoughts is to spark future reflections and discussions over the challenges in annotating stuttered speech. Please refer to our supplementary files for a more detailed list and related example clips.

- (1) Lack of Embodied Understandings of Stuttering
 - Annotators didn’t differentiate stuttering events and ‘natural’ or intentional disfluencies.
 - Annotators didn’t understand the labels as how PWS feel about them.
 - Speakers’ internal experience of their speech often differs from what listeners are able to perceive.
- (2) Diverse and Intersecting Stuttering Patterns
 - Hard to objectify stuttering events because of the diverse speech patterns, e.g., many instances resemble mixed events, PWS may draw on techniques to manage stuttering.
- (3) Challenges in Transcribing Disfluent Speech Verbatim
 - Repeat fast and repeat many times, e.g., [n-n-n/p-n-]/snavigate to mom’s house.
 - Sound repetition vs. word repetition for single-syllable words, e.g., [Aa]/s [add]/r add a pocket projector for Katie to my gift list.
- (4) Inherent Subjectivity in Speech Perception
 - The perception of elongation of vowels is inherently subjective especially without listening to the speech flow, and for long vowel and compound vowel: *hear, been, language, oh, ends, two, so, okay, wow*.
- (5) Distortions Introduced by Audio Segmentation
 - Hard to contextualize the sounds because of the audio segmentation.

| Stutter type | Code | Example | Notes |
|------------------------|-------|--|--|
| Block | /b | - My /bname (blocking on “n”) - Spa/bghetti (blocking on “g”) | A blocking pause before or within a word. Insert /b right before the next sound/word. While not being prescriptive, to attend to these blocks, you can pay attention to the airflow, a gush of air right before a sound (e.g. a strong airy “t” sound in “turn”) can indicate a block, even when there is no clear pause in speech. |
| Prolongation | /p | M/pommy Mommy, with “m” sound elongated. | Elongated syllable. Insert /p right after the prolonged syllable |
| Sound repetition | []/s | [pr-pr-pr-]/sprepare Prepare, with “pr” repeated three times. Verbatim transcript: “pr-pr-pr-prepare” | Repeated syllables. Transcribe the exact number of times the sound/syllable is repeated, put a “-” after each one and put all extra text in brackets, insert the /s right after the brackets. |
| Word/Phrase repetition | []/r | [my my]/r my name | The same word or phrase is repeated. Transcribe the exact words/phrases as repeated, put the extra parts in brackets with space in between each word, insert /r right after the bracket, leave a space between /r and the next word. |
| Interjection | []/i | I [uh]/i work | Common filler words such as “um” or “uh” or person-specific filler words that individuals use to cope with their stutter (e.g., some users frequently say “you know” as a filler). Transcribe the stuttering-related filler words as they are and put them in the brackets, followed by the /i mark. Leave a space between /i and the next word. * You need to transcribe non-stuttering filler words verbatims well, but do NOT use []/i to mark them as stuttering events. |

Table 7: Basic stutter types and corresponding annotation markups

- Hard to tell the words because of audio segmentation, e.g., *um* vs. *I’m*, and vs. *uh*.
- (6) Impact of Audio Processing Methods
- Easy to miss non-speech content because of how Pratt renders segments; annotators might click the segment while missing cues outside of the segments.

C Interview Guide Used in Formative Studies

- (1) How would you describe stuttering to people who do not stutter?
- (2) What would you want AI models to learn about stuttering?
- (3) Can you walk me through a typical workflow you followed when annotating the speech data?
- (4) What differences did you observe between the labels you assigned and those from crowdworkers?
 - How did you feel about these differences?
 - What factors do you think might cause these differences?
- (5) I picked some examples from the dataset you annotated. Can you walk me through how you did annotations for these examples?
 - Did you feel anything hard to make a decision during the process?

- Do you remember any other hard cases? What makes these examples hard?
- (6) If you have an opportunity to re-design the annotation process, how would you design the annotation task?
- What labels/features would you design to represent stuttering?
 - What expertise do you think PWS or non-PWS have in annotations?