Towards Fair and Inclusive Speech Recognition for Stuttering: Community-led Chinese Stuttered Speech Dataset Creation and Benchmarking

QISHENG LI, AImpower.org, USA
SHAOMEI WU, AImpower.org, USA

Despite the wide adoption of Automatic Speech Recognition (ASR) models in voice-operated products and conversational AI agents, current ASR models perform poorly for people who stutter (PWS). One source of the performance disparity is the lack of ample, representative data of stuttered speech in the development of ASR models. To fill the gap, we present the first stuttered speech dataset in Mandarin Chinese, created by a grassroots community of Chinese-speaking PWS to facilitate the development of inclusive and fair speech AI. Collected from 70 speakers with a wide range of stuttering characteristics, this dataset contains speech samples of both spontaneous conversations and voice command dictations from each speaker. Our analysis of the dataset shows the diversity and variability of stutters captured, highlighting its unique value in authentically representing the stuttering community in AI data. Leveraging this dataset, we benchmark popular ASR models and uncover their embedded biases for speech fluency.

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

Additional Key Words and Phrases: AI FATE, datasets, benchmark, speech technology, accessibility, stuttering, stuttered speech

ACM Reference Format:
Qisheng Li and Shaomei Wu. 2024. Towards Fair and Inclusive Speech Recognition for Stuttering: Community-led Chinese Stuttered Speech Dataset Creation and Benchmarking. In . ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Stuttering affects approximately 1% of the population worldwide [4]. Although the condition is typically characterized by the speech behaviors such as repetitions ("li-li-like this"), prolongations ("lllllike this"), and blocks ("l—ike this"), people who stutter (PWS) often experience significant structural disadvantages beyond these observable "speech disfluencies". Ample research has demonstrated that people who stutter (PWS) regularly face negative listener reactions, such as stigma [6] and discrimination [8], which can affect all aspects of life, including establishing social and romantic relationships [5, 11], achieving educational goals [14], and pursuing employment opportunities [16].

As AI-powered Automatic Speech Recognition (ASR) systems becoming a ubiquitous part of today's communication ecosystem, research has shown that ASRs have significant difficulty in decoding stuttered speech, resulting in three to four times higher word error rate (WER) compared to non-stuttered speech [20]. Often, ASR-powered systems misinterpret the speech of PWS, cut them off while they are speaking, or simply be unable to provide valid responses [3, 20]. The inability of ASR systems to work with stuttered speech not only creates additional barriers for PWS to interact with popular products and services like smart speakers, in-car navigation systems, and automatic phone menus, but also leads to potential psychological harms such as heightened self-consciousness and reduced self-esteem [8, 35].

Manuscript submitted to ACM

The lack of adequate stuttered speech dataset has been a key bottleneck for addressing ASR performance disparities [21, 22]. Existing stuttered speech datasets, such as FluencyBank [28] and LibriStutter [18], are often limited in size, representativesness, and annotation consistency [21]. Recent efforts in collecting atypical speech for inclusive ASR also face challenges in engaging and authentically representing the stuttering community [22]. Distinct from other speech etiologies such as Amyotrophic Lateral Sclerosis (ALS), stuttering is highly variable and inherently social: the severity and patterns of stuttering can vary significantly across individuals, the situations, and the conversation partners [9, 32]. As a result, conventional speech data collection approach that records participants' monologue responses to given prompts [2, 22, 25] often works poorly at capturing authentic, real-world stuttering patterns.

To fill in this gap, we present ***StammerTalk Chinese Stuttered Speech Dataset***, the first and largest stuttered speech dataset in Mandarin Chinese, created by StammerTalk - a grassroots community of Chinese speaking people who stutter. Including about 50 hours of speech with both spontaneous conversations and voice command dictations from 72 individuals who stutter, the StammerTalk dataset offer unprecedented versatility and authenticity comparing to existing stuttered speech datasets. Through descriptive analysis of the dataset, as well as the evaluation of state-of-the-art ASR models using this data, our work shows the unique value of community-created stuttered speech dataset at capturing the heterogeneity and variability of stuttering, highlighting its efficacy at uncovering fairness issues in existing ASR models and raising awareness of the needs of PWS in the AI community. Our work contributes to existing literature on ASR fairness, inclusivity, and equity, advocating for open and respectful collaboration between the AI and the disability communities in the full cycle of speech AI design and development.

2 RELATED WORK

2.1 Stuttering

Stuttering is a neurodevelopmental condition that impacts PWS in behavioral, emotional, and cognitive aspects [4]. Besides speech struggles, PWS often develop adverse emotional and cognitive reactions of stuttering, such as fear, guilt, shame, and self censorship. The behavior, emotional, and cognitive components of stuttering make it highly variable and diverse [4, 9, 32]. Across individuals who stutter, a wide range of speech characteristics exist, from observable speech disfluencies like word repetitions, to "masked" disfluencies such as word switching and circumlocution [9]. Even for the same speaker, their stuttering pattern and frequency can vary from almost complete fluency when speaking to oneself alone, to severe disfluencies when speaking over the phone or to a group [32]. Such inherent irregularities in stuttered speech make it particularly challenging when applying modeling techniques that rely on statistical methods and pattern recognitions – a dominant approach in modern ASR. As a result, today's ASR systems have great difficulty in understanding and interacting with users who stutter [3], showing as high as 50% WER for severely stuttered speech, 10 times the reported consumer average of 5% [20]. Our work contributes to a deeper understanding of AI FATE issues for people who stutter, a population profoundly affected yet often overlooked by speech AI technologies.

2.2 Inclusive ASR for Stuttering

In the domain of ASR, ableist assumptions about the syntax and temporality of human speech are prevalent. For example, endpointer models – a component of ASR that identifies the end of a utterance – frequently truncate the speech input from PWS as these models have learned auditory features (including the duration of silence between sounds) from fluent speech [20]. Similarly, ASR decoders tend to inject unrelated, real words in place of partial or whole word repetitions [20, 26]. Recent technical explorations showed promise in addressing these biases by fine-tuning ASR decoders with disfluent speech [20, 26], adjusting endpointer thresholds [20], and model personalization [15, 33]. In parallel, research on stuttering event detection [1, 18] explores a two-step "detect and pass" approach to help ASR models better process stuttered speech [30].

However, current solutions often require a relatively large amount of annotated personal speech data [15, 33], with improvements limited to short phrases [15], prompted speech [15], pre-defined vocabularies and scenarios [20, 26, 33], and speech with reduced articulations (e.g. speech by patients with ALS, CP, or Pakinson's Disease) [15, 33]. The question remains about their generalizability to everyday interactions with ASR-powered systems by PWS "in the wild". Overall, there is still a significant gap in speech technology accessibility for people who stutter. Our work intends to expand current technical efforts to address this problem by auditing popular ASR models, using stuttered speech in different scenarios and from a diverse set of speakers with varying stuttering patterns, to pinpoint existing pain points and uncover new opportunities for improvement.

2.3 Stuttered Speech Datasets

Recognizing the need for diverse speech data in the development of more inclusive and robust ASR models [23, 26, 38], there have seen several industry-led efforts to collect more diverse speech samples across languages [2], accents [2, 25], and speech disabilities [22, 34]. Despite these efforts, public and high-quality stuttered speech data remain scarce.

Considered highly sensitive personal data, stuttered speech collected and annotated by companies are often inaccessible to the broader research community [20, 22, 26, 30, 38]. Datasets collected by academic researchers, such as FluencyBank [28], the University College London's Archive of Stuttered Speech (UCLASS)[17], were limited in sizes and annotation consistency as they were intended as therapy resources [21]. Open and scalable datasets, such as SEP-28k [21] and LibriStutter[18], also have various shortcomings in terms of annotation completeness, representativeness, and authenticity. Collected from public podcasts by people who stutter, SEP-28k consists of 28K 3-sec audio clips labeled with only stuttering events (i.e. whether the clip contains prolongation, sound repetition, etc.) but no text transcriptions [21]. While LibriStutter does provide transcriptions, it contains no real speech from PWS but synthetic stuttering utterances (repetitions, prolongations, interjections) injected into prompted speech extracted from audio books [18].

By introducing StammerTalk dataset, the first and largest corpus of stuttered speech in Mandarin Chinese, our work contributes to the ongoing efforts in representing stuttered speech and the stuttering community in AI data. We also highlight the value of community-driven data collection and annotation process to better capture the authentic and diverse expression of stuttering and represent the voice of the community in AI speech data development.

3 STAMMERTALK DATASET

The dataset was created by StammerTalk (口吃说) community members, through remote recording sessions via videoconferencing. Each session was conducted by a StammerTalk volunteers, who also stutter, with a participant who stutters. The recorded speech contains both unscripted conversations between the volunteer and the participant, and the dictation of a list of 200 voice commands by the participant. 70 adults who stutter (AWS) participated in the recording with two StammerTalk volunteers, resulting in a dataset of 50 hours speech from 72 AWS. The recorded speech was transcribed verbatim, with five distinct stuttering event annotations (**[]** - Word-level repetition, **/r** - sound repetition, **/b** - blocks, **/p** - prolongation, **/i** - interjection) embedded in markups. See Appendix A for more details on data collection and annotation, as well as the participants recruitment and selection process.

4 DESCRIPTIVE ANALYSIS

In this section, we provide descriptive statistics about the StammerTalk dataset, showcasing its scale and data diversity, to illustrate its unique value in providing an authentic and comprehensive representation of stuttering and stuttered speech for ASR models and the ASR research community.

4.1 Scale and Scope

We describe the scale and scope of the StammerTalk dataset in terms of speakers, speech duration, stuttering events, and speech content. Key statistics for these aspects are provided in Table 1, along with existing datasets for comparison.

Table 1. **Dataset scale and scope as characterized by speech duration (Duration), the number and types of speakers (Speakers), whether it provides speech transcription (Transcription), types of speaking tasks (Tasks), and Language.**

Dataset	Duration	Speakers	Transcription	Tasks	Language
FluencyBank* [28]	3.5 hrs	32 AWS	Yes	conversation, reading article	English
LibriStutter [19]	20 hrs	50 non-PWS	Yes**	audiobook	English
UCLASS* [17]	53 mins	25 CWS	Yes	conversation	English
SEP-28k [21]	23 hrs	not reported	No	podcast	English
StammerTalk	**50 hrs**	**72 AWS*******	**Yes (verbatim)**	**conversation, voice commands**	**Chinese**

* Limited to the transcribed portion of the dataset.
** Stuttered utterances are masked in the transcription as "STUTTER".
*** Including two StammerTalk volunteers.
Abbreviations: AWS - adults who stutter; CWS - children who stutter; PWS - people who stutter.

Duration. A total duration of 50-hours speech data were included in StammerTalk dataset from 70 data collection sessions. Excluding the StammerTalk volunteers, each participant contributed on average 33.0 minutes of total speech (*min*=17.2, *max*=49.9, SD=7.32), with an average of 17.8 minutes (*min*=7.24, *max*=34.93, SD=5.6) for conversations and an average of 15.23 minutes (*min*=6.45, *max*=27.6, SD=5.23) for voice command dictation. Many participants found speaking with another PWS both rare and pleasant [12], thus spent more time on the conversations.

Stuttering Events. Excluding stuttering event annotations, the text transcription of StammerTalk dataset contains 429K Chinese characters (275K for conversations, 174K for voice command dictation), with in total 28,310 stuttering events annotated. Table 2 presents the frequency and distribution of annotated stuttering events in the StammerTalk dataset, in comparison with existing stuttered speech datasets that have stuttering event annotation, highlighting the quantity and diversity of stuttering events captured in our data. In particular, we compute the *Average Stuttering Rate* by dividing the total count of stuttering events by the duration of the speech, and found that conversational speech in the StammerTalk dataset exhibits approximately 25% more frequent stuttering occurrences compared to the stuttering-related podcast (SEP-28k) and synthetic stuttering speech (LibriStutter). In terms of the distribution of different stuttering events, we also observed a significant shift towards more word and phrases repetitions and less sound repetitions in the StammerTalk dataset, signaling potential phonological differences between stuttering in Chinese and in English. Meanwhile we notice that SEP-28k dataset contains almost twice more interjections than in StammerTalk conversations, which could be attributed to different definitions of interjections in these two datasets. While SEP-28k considers any filler words - such as "um," "uh," and "you know" - as stuttering interjections, StammerTalk's annotation excludes natural interjections that blend into the speech flow. In general, the distributional differences in stuttering types not only reflect the innate linguistic and situational differences between existing datasets and StammerTalk dataset, but also highlights the challenges to identity and understand stuttering from an observer's perspectives - an increasingly important issue for fluency researchers and speech language pathologists [37].

Table 2. **Overall frequency and distribution of annotated stuttering events**

	Avg. Stuttering Rate (per minute)	Total Stuttering Events	Event Type Distribution				
			[]	/b	/p	/r	/i
StammerTalk: Conversation	15.83	19,674	42%	6%	18%	9%	25%
StammerTalk: Command	4.22	8,636	53%	8%	17%	16%	6%
SEP-28k [21]	12.26	17,273	23%	28%	23%	19%	49%
LibriStutter [18]	12.5*	15,000*	20%	20%	20%	20%	20%

* Stuttering utterances were synthetically generated.

Content. The StammerTalk dataset contains both voice command dictation and unscripted conversations in Chinese, with topics around stuttering and voice products. The content of the speech thus become a resource for community advocacy and outreach, raising public awareness and understanding on a largely invisible issue. As the only stuttered speech corpora in a non-European language to the best of our knowledge, the StammerTalk dataset fills in an important language gap for stuttered speech and opens a door to quantify the linguistic and cultural differences in stuttering in Chinese versus other - mostly Eurocentric - languages.

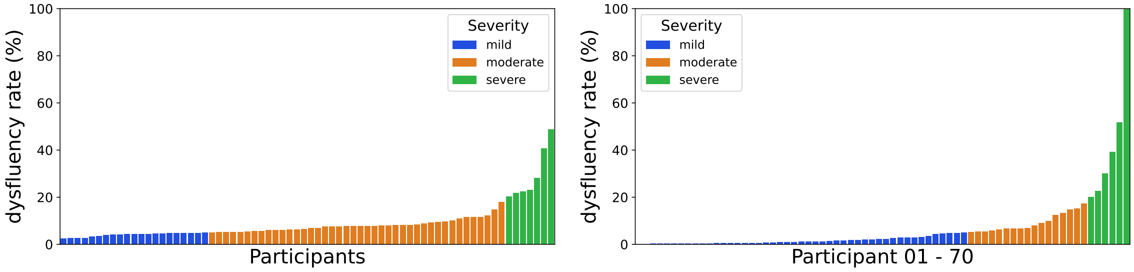
To summarize, as illustrated in Table 1, the StammerTalk dataset surpasses existing datasets in its scale and speaker pool. It contains 20 times more transcribed speech samples from people who stutter than what is available today (i.e. FluencyBank), and includes two to three times more stuttering speakers. It also provides both stuttering event annotations and verbatim transcription, allowing its versatile applications for a wide range of technical approaches. Instead of podcasts or audio books, the StammerTalk dataset contains unscripted conversations and voice commands that are closer to real-world speech product use cases such as meeting transcriptions and speech-operated devices.

4.2 Speech Diversity

Stuttering is not a monolith. The frequency and distribution of stutters also vary significantly across participants and speaking tasks - a common source of insecurity and frustration for PWS [32]. Here we describe the variability and heterogeneity of stuttered speech, as reflected in the StammerTalk dataset.

Stuttering frequency. While all participants self-identified as people who stutter, their frequency of stuttering varies. To quantify individual stuttering frequency, we calculate *dysfluency rate*, as defined in [13, 20], by dividing the total number of stuttering events over the total transcribed character count in the speech of each speaker. Following conventional approaches [13, 20], participants are categorized based on *dysfluency rate* into three groups, corresponding to mild (0-5% of words with dysfluencies), moderate (6-20% of words), and severe (over 20% of words) stuttering. As illustrated in Fig. 1, while the participants stutter more in Conversations (mean=9%) than in Command Dictation (mean=7.1%), the frequency of stuttering varies more in Command Dictation (std=0.15) than in Conversation (std=0.08). As a result, the grouping of speakers varies across two tasks: for Dictation, 47, 17, and 6 speakers are categorized as mild, moderate, and severe stuttering, respectively, whereas for Conversation, the numbers are 21, 42, and 7.

The variation in *dysfluency rates* across different tasks and speakers reflects the heterogeneity within the stuttering community. For some, reading is much easier than spontaneous conversations; whereas for others, reading could be



(a) Individual dysfluency rates in StammerTalk Conversation

(b) Individual dysfluency rates in StammerTalk Dictation

Fig. 1. **Dysfluency rates of all 70 participants in Conversation and Dictation tasks, sorted from low to high, and categorized into mild (0-5%), moderate (5-20%), and severe (20%+) stuttering groups.**

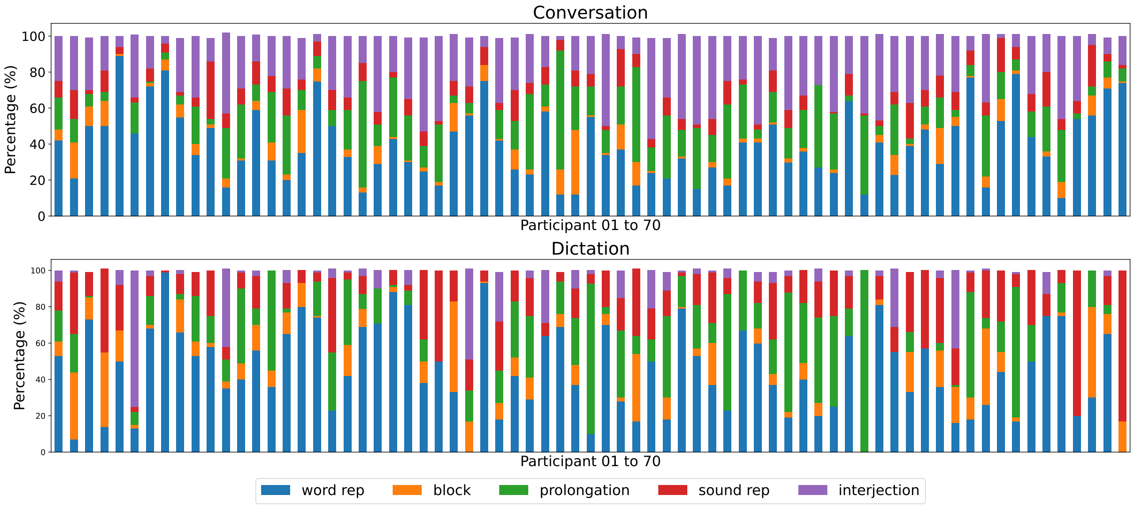


Fig. 2. **Breakdowns of five annotated stuttering events for 70 participants**

extremely challenging (100% *dysfluency rate*). The StammerTalk dataset highlights the dynamic and situated nature of stuttering: its severity varies not just among individuals but also within the same individual in different scenarios.

Stuttering variability. PWS often stutter differently: some might show more repetitions, some with more blocks, while some stutter covertly [9, 32]. Fig 2 shows the breakdowns of annotated stuttering events, for all 70 participants, highlighting the diversity in their stuttering patterns. It also illustrates the change in stuttering patterns for the same speaker with different tasks: participants often have relatively more interjections in conversations, but show an increase in sound repetitions when dictating commands.

Contrasting to previous stuttered speech datasets [18, 21], the StammerTalk dataset captures a wide spectrum of stuttering frequency and patterns across PWS in different scenario, providing a much more authentic and comprehensive representation of the stuttering community for AI models.

5 BENCHMARKING ASR MODELS WITH STAMMERTALK DATASET

Leveraging the StammerTalk dataset, we audit two popular opensourced ASR models: Whisper (large-v3)[1] and wav2vec2.0 (large-chinese-zh-cn)[2]. Our initial audit is intended to benchmark state-of-the-art ASR model performance on Chinese stuttered speech and to inform the areas for further development.

We measure the character error rate (CER), substitution (SUB), insertion (INS) and deletion (DEL) error rates of both models on each severity level. Character Error Rate (CER) measures the performance of ASR systems for Mandarin speech, by counting the errors made by the model at the character level, including substitutions (SUB), insertions (INS), and deletions (DEL), of characters comparing to the reference. We compare ASR outputs with two types of ground truth transcriptions: 1) a **semantic** transcription with word repetitions and interjections excluded; 2) a **literal** transcription with the stuttered utterances kept verbatim. We remove all stuttering event annotations in both cases.

The benchmarking results with Whisper model are shown in Fig. 3. We notice that it performs reasonably well on the mildly stuttered speech, achieving a CER of 10.74% for unscripted conversation (Fig. 3a). However, CER increases with stuttering severity, reaching 13.46% for the moderately and 19.46% for the severely stuttered speech. These CERs are significantly highers than the baseline results reported for the general population for the Whisper model on Mandarin speech, namely, 12.8% on the Common Voice 15 dataset and 7.7% on the FLEURS dataset[3].

Comparing Fig. 3a with Fig. 3b, we find a sharp increase in deletion errors (DEL) when referencing on literal transcriptions. Further inspection of the results shows that Whisper has difficulties in generating disfluent literal transcriptions, often "smoothening" its transcriptions by removing repeated words or phrases. We provide examples in Table 3 in Appendix B to illustrate this behavior. As presented in Fig. 3c and Fig. 3d, we find that CERs are higher for Dictation tasks compared to Conversation. We suspect that Whisper rely on language model to generate reasonable transcription within a semantic context, thus could perform sub-optimally for shorter speech, such as voice commands. Wav2vec model, in contrast, performs 1.5 to 2 times worse than Whisper, and produce a lot more substitution mistakes. Manual inspection finds that wav2vec model often substitutes a character with its homophones, undervaluing the semantic cohesiveness or sentence fluency. More detailed results for wav2vec model can be found in Appendix B.

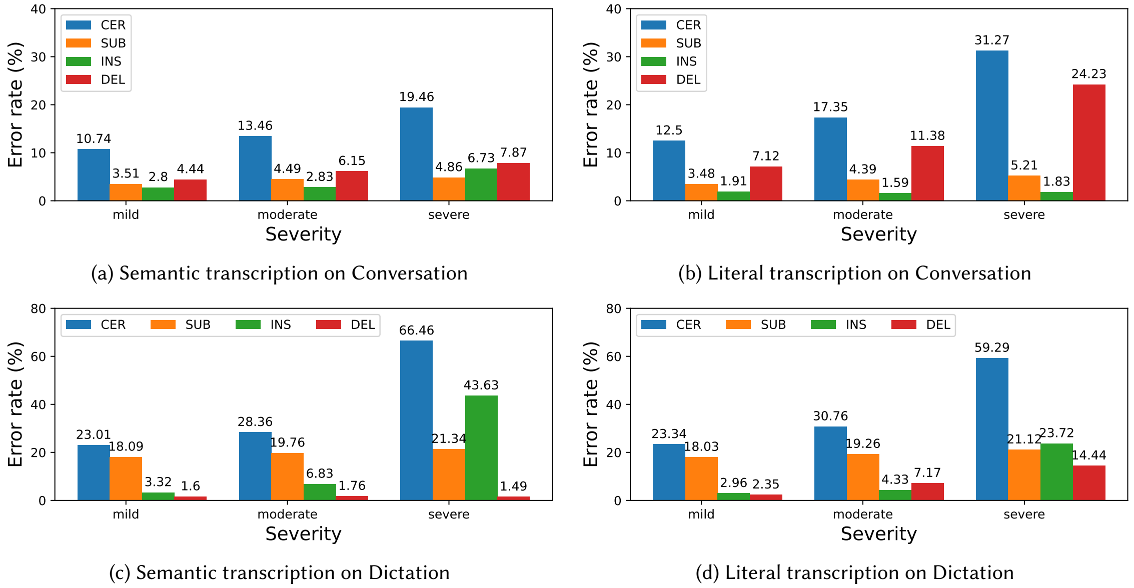
6 DISCUSSION AND FUTURE WORK

Representing the disability community adequately and authentically in AI data has been a prominent challenge in AI fairness and accessibility [24, 27]. This challenge is even more pronounced for stuttering, an "invisible" disability that is highly variable, social, and situational. Collected and curated by StammerTalk community, the StammerTalk dataset surpasses existing stuttered speech datasets in its technical and informational values.

The unprecedented size and diversity of stuttered speech in the StammerTalk dataset allows for a wide range of technical explorations and innovations. For example, as the divergent stuttering patterns captured for Voice Command Dictation and Natural Conversation accentuates the importance of situational context in understanding stuttered speech, future ASR models should anticipate and incorporate such situational differences, using features exacted from the StammerTalk dataset as a starting point.

Additionally, the rigorous verbatim transcription, annotated with specific stuttering events, create opportunities for ASR systems to recognize and transcribe stuttered utterances truthfully, which not only provides a more accurate transcription but also openly accepts stuttering in human communications - an attitude that is clinically proved to

[1] https://github.com/openai/whisper
[2] A fine-tuned version of wav2vec2.0 optimized for Mandarin speech, seehttps://huggingface.co/wbbbbb/wav2vec2-large-chinese-zh-cn
[3] https://github.com/openai/whisper?tab=readme-ov-file



(a) Semantic transcription on Conversation

(b) Literal transcription on Conversation

(c) Semantic transcription on Dictation

(d) Literal transcription on Dictation

Fig. 3. **Character error rate (CER), substitution (SUB), insertion (INS) and deletion (DEL) error rates for Whisper evaluated on StammerTalk Conversation and Dictation.**

benefit PWS in a long term [31]. In contrast, our audit of OpenAI's Whisper model finds a forced "smoothening" of stuttered speech in its transcriptions, uncovering its embedded ableist biases that value fluency over dysfluency [31].

Despite StammerTalk dataset's unprecedented scale, stuttered speech is still immensely under-represented in ASR (e.g. Whisper was trained over 680,000 hours of multilingual data). Future work could further expand the scale of high-quality stuttered speech dataset by replicating StammerTalk's data collection model in more languages and dialects. On the other hand, while existing work on fine-tuning general ASR model with a small amount of stuttered speech show promise [20, 26], future work could leverage the StammerTalk dataset to further explore this direction.

Future work is also required in managing the StammerTalk dataset. Although the community intend to opensource their data for scientific and technological advancements, they need to navigate complex legal and technical systems to identify the suitable infrastructure for collective ownership, personal data protection, and cross-border data regulations.

Our initial benchmarking takes a critical step in understanding the interaction between existing ARS models and stuttered speech. However, our results are preliminary. Future work should include a wider set of ASR models and investigate recognition errors more thoroughly and in-depth.

7 CONCLUSION

In conclusion, the rise of speech AI technologies, while benefiting many, has introduced additional accessibility barriers for people who stutter. Our research tackles the performance disparity in modern ASR systems for stuttered speech by introducing a community-created, large scale stuttered speech corpus in Chinese, and demonstrating its effectiveness in benchmarking and diagnosing state-of-the-art ASR models for stuttered speech. Our analysis of the StammerTalk dataset demonstrate the scope and diversity of stutters it captured, highlighting its unique value for authentically representing the stuttering community in ASR data. Our audit of leading ASR models using the StammerTalk dataset confirmed the performance regressions observed in other languages [20], establishing a much needed baseline to bootstrap the progress in improving ASR models for Chinese stuttered speech. Our investigations into the errors made by audited ASR models shed light on the ableist assumptions currently built-in in advanced ASR models, uncovering opportunities for developing fair and inclusive speech technologies for people who stutter.

REFERENCES

[1] Sadeen Alharbi, Madina Hasan, Anthony JH Simons, Shelagh Brumfitt, and Phil Green. 2018. A lightly supervised approach to detect stuttering in children's speech. In *Proceedings of Interspeech 2018*. ISCA, 3433–3437.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4218–4222. https://aclanthology.org/2020.lrec-1.520

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

[4] O. Bloodstein, N.B. Ratner, and S.B. Brundage. 2021. *A Handbook on Stuttering, Seventh Edition*. Plural Publishing, Incorporated. https://books.google.com/books?id=Abw0EAAAQBAJ

[5] John Van Borsel, Marie Brepoels, and Janne De Coene. 2011. Stuttering, attractiveness, and romantic relationships: The perception of adolescents and young adults. *Journal of Fluency Disorders* 36, 1 (2011), 41–50.

[6] Michael P. Boyle. 2018. Enacted stigma and felt stigma experienced by adults who stutter. *Journal of Communication Disorders* 73 (2018), 50–61. https://doi.org/10.1016/j.jcomdis.2018.03.004

[7] Courtney T Byrd, Zoi Gkalitsiou, Joe Donaher, and Erin Stergiou. 2016. The client's perspective on voluntary stuttering. *American Journal of Speech-Language Pathology* 25, 3 (2016), 290–305.

[8] Geoffrey A. Coalson, Alexus Crawford, Shanley B. Treleaven, Courtney T. Byrd, Lauren Davis, Lillian Dang, Jillian Edgerly, and Alison Turk. 2022. Microaggression and the adult stuttering experience. *Journal of Communication Disorders* 95 (2022), 106180. https://doi.org/10.1016/j.jcomdis.2021.106180

[9] 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

[10] C. D. Constantino, W. H. Manning, and S. N. Nordstrom. 2017. Rethinking covert stuttering. *Journal of Fluency Disorders* 53 (2017), 26–40.

[11] S. Davis, P. Howell, and F. Cooke. 2002. Sociodynamic relationships between children who stutter and their non-stuttering classmates. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 43, 7 (2002), 939–947.

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

[13] Andrews Gavin and MARY HARRIS. 1964. The Syndrome of Stuttering'. *Clinics in developmental Medicine* 17 (1964).

[14] Hope Gerlach-Houck, Kristel Kubart, and Eilidh Cage. 2023. Concealing Stuttering at School: &#x201c;When You Can't Fix It&#x2026;the Only Alternative Is to Hide It&#x201d;. *Language, Speech, and Hearing Services in Schools* 54, 1 (2023), 96–113. https://doi.org/10.1044/2022_LSHSS-22-00029 arXiv:https://pubs.asha.org/doi/pdf/10.1044/2022$_{L}SHSS - 22 - 00029$

[15] Jordan R Green, Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, et al. 2021. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases.. In *Interspeech*. 4778–4782.

[16] Gerlach H, Totty E, Subramanian A, and Zebrowski P. 2018. Stuttering and Labor Market Outcomes in the United States. *J Speech Lang Hear Res.* 61, 7 (2018), 1649–1663. https://doi.org/10.1044/2018_JSLHR-S-17-0353 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6195060/.

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

[18] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2021. FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (sep 2021), 2986–2999. https://doi.org/10.1109/TASLP.2021.3110146

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

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

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

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

[23] Valentin Mendelev, Tina Raissi, Guglielmo Camporese, and Manuel Giollo. 2021. Improved robustness to disfluencies in rnn-transducer based speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6878–6882.

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

[25] Meta. 2023. Speech Fairness Dataset. https://ai.meta.com/datasets/speech-fairness-dataset/.

[26] Vikramjit Mitra, Zifang Huang, Colin Lea, Lauren Tooley, Panayiotis Georgiou, Sachin Kajarekar, and Jefferey Bigham. 2021. Analysis and Tuning of a Voice Assistant System for Dysfluent Speech. https://arxiv.org/pdf/2106.11759.pdf

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

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

[29] J.G. Sheehan. 1970. *Stuttering: Research and Therapy*. Harper & Row. https://books.google.com/books?id=0KhrAAAAMAAJ

[30] Olabanji Shonibare, Xiaosu Tong, and Venkatesh Ravichandran. 2022. Enhancing ASR for Stuttered Speech with Limited Data Using Detect and Pass. arXiv:2202.05396 [eess.AS]

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

[32] 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.

[33] Jimmy Tobin and Katrin Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6637–6641.

[34] UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN. 2023. Speech Accessibility Project. https://speechaccessibilityproject.beckman.illinois.edu/. Accessed on 1/21/2024..

[35] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 109, 14 pages. https://doi.org/10.1145/3544548.3581357

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

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

[38] 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).

A DATASET CREATION

A.1 Date Collection

The data collection was conceptualized and executed by StammerTalk (口吃说), a grassroots online community for Chinese-speaking people who stutter. Frustrated by the poor performance of automatic speech recognition (ASR) systems for stuttered speech, the StammerTalk community self-organized to create and curate the first and largest Chinese stuttered speech corpus to improve their experience with speech interfaces and products.

The data collection process spanned over nearly one year, from January 2023 to November 2023. Two volunteers from the StammerTalk community recruited participants within the community, and conducted 60-minute data collection sessions with each participant remotely via Zoom or Tencent Meet. Each session is structured as 1) an 5-min introduction; 2) a 30-min **Unscripted Spontaneous Conversation** around participant's personal background, interests, and lived experiences with stuttering; and 3) a 30-min **Voice Command Dictation** in which the participant was provided a list of 200 common voice commands to read them out in order. The participant would be asked to voluntarily stutter [7] and/or read the list repeatedly, if they read too fluently or too fast. The latter two components of the session were audio recorded as raw speech samples, which were later annotated and transcribed for analysis and training ASR models. Approximately an hour of speech data was collected from each session.

A.2 Data Annotation

Pro bono speech data annotation service was provided by a commercial speech annotation company in China (name redacted for anonymity). Given the absence of guidelines for annotating stuttered speech in Chinese, StammerTalk volunteers extended existing annotation guidelines for fluent speech with stutter-specific instructions adopted from similar efforts in English [21]. The detailed guidelines were developed with input and feedback from speech and language pathologists (SLPs), fluency researchers, AI researchers, and other PWS, and refined over three iterations with two professional speech annotators who do not stutter. Feedback and training were provided by StammerTalk volunteers to the annotators to successfully identity, annotate, and transcribe stuttered utterances. The trained annotators then performed the speech-to-text transcription and stuttering event annotation for all recorded speech data through Praat[4], providing phonetic-level annotation and transcription of stuttered speech. As a result, the transcriptions are easily consumable by machines.

Similar to previous work [18, 21], five types of stutters were specified by the annotation guidelines, including:

- **[]**: **Word-level repetition**. Repeated words or phrases.
- **/r**: **sound repetition**. Repeated sounds, such as a consonant or vowel, that do not constitute an entire word.
- **/b**: **blocks**. Prolonged blocks or unnatural silence.
- **/p**: **prolongation**. Prolonged phonemes.
- **/i**: **interjection**. Excessive utterances like '嗯' (hmm), '啊' (ah), or '呃' (um). Notably, natural sounding interjections that do not disrupt the speech flow are excluded from this category.

The StammerTalk community made the deliberate decision to have **the transcription performed verbatim, with stuttering events annotations embedded in markups**. While existing stuttered speech datasets often chose to mask stuttered utterances in the transcription (e.g. LibriStutter [18]), the StammerTalk community chose to represent the stuttered utterances authentically and explicitly in the transcriptions, as it not only provides a more accurate

[4] Pratt is a commonly used software package for speech phonetics analysis. See https://www.fon.hum.uva.nl/praat/.

transcription but also reaffirm the space for stuttering in human speech and for AI models [31]. An example annotation is provided below. Note that the same word can have multiple stuttering events, such as both block and prolongation.

Example Annotation: 我叫[我叫/p]小/b明，我[我我]住/p/b在呃/i北/r京

Interpretation: I am I am (multi-words repetitions and prolongation of "am") Xiao (block) Ming, I I I (single word repetition) live (prolongation and block) in um (interjection) Bei ("b" sound repetition) Jing.

A.3 Participants

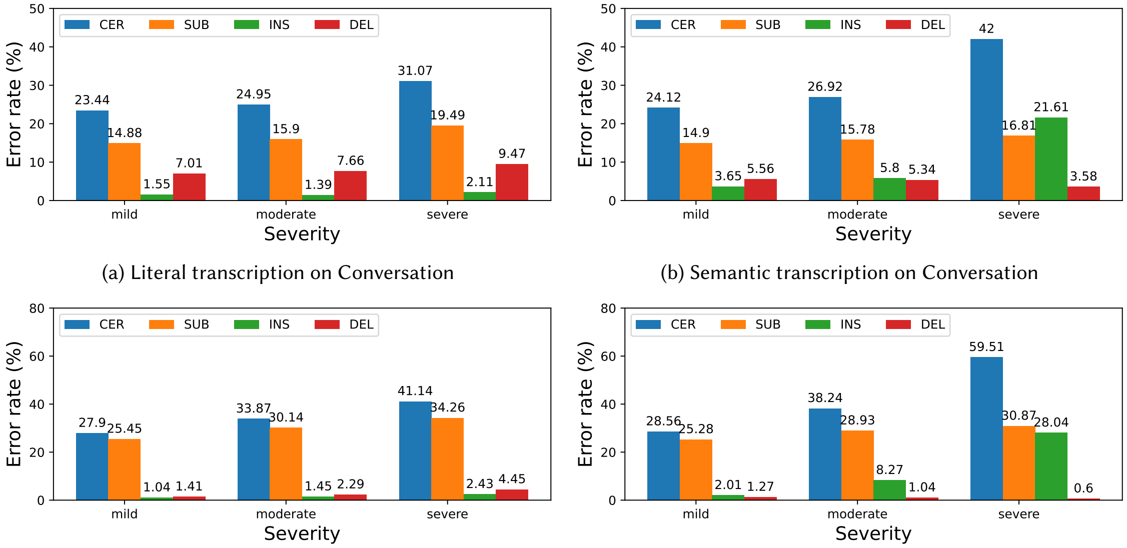
70 people who stutter (not including the StammerTalk volunteers) had participated in the data collection process, contributing approximately 70 hours of speech to the dataset. The participants were recruited through a recruitment message posted by StammerTalk's public account on WeChat,and compensated with a monetary compensation of ¥100 RMB cash via WeChat pay, as well as a swag by the commercial speech annotation company.

Although the participants were asked to complete the Overall Assessment of the Speaker's Experience of Stuttering (OASES) [37], it was not used as a selection criterion but a resource for the participants. This recruitment method was design to accommodate the heterogeneity within the stuttering community and "hidden" nature of stuttering [29]: anyone who *self identifies* as a person who stutter is welcomed, including those who stutter covertly[5].

After completing the data collection session, the participants were asked to complete an exit survey to provide their demographics, contact information, and general sentiment about the data collection session. Although optional, the survey was completed by 95% of the participants, showing highly positive experience (in a 1-5 scale, 70% "5 - extremely satisfied", 100% above 3), a sharp contrast with the heightened stress and "performance anxiety" observed in previous research [27, 36]. This result highlights the emotional benefits of community-led stuttered speech collection beyond the tangible dataset. Most participants (64 out of 70) are from mainland China, and 24 of them self-reported as female.

[5] Covert stuttering is a type of stuttering with little or no disfluencies that can be effectively passed as fluent speech to the listener [10].

B ADDITIONAL BENCHMARKING RESULTS



(a) Literal transcription on Conversation

(b) Semantic transcription on Conversation

(c) Literal transcription on Dictation

(d) Semantic transcription on Dictation

Fig. 4. **Character error rate (CER), substitution (SUB), insertion (INS) and deletion (DEL) error rates for fine-tuned Wav2Vec 2.0 evaluated on StammerTalk Dictation**

Table 3. **Whisper model "smooths" the transcriptions by removing words with low semantic value, as indicated by the underlined characters.**

Reference	Whisper model hypothesis
就在那个继续深造的也有	就在继续深造的也有
嗯我觉得深圳他到处他都是花钱的地方就是吃喝玩乐他肯定是	我觉得深圳到处都是花田的地方吃喝玩乐肯定是

Table 4. **Three examples of utterances from a severe PWS, characterized by frequent word repetitions. The Whisper model produced insertion errors for these repetitions, as indicated by the underlined characters.**

Annotation	Whisper model hypothesis
经/r/b济产业啥[啥啥]嗯/i是[是]有关[关关/r/b关/b关/b关/b]系的	经济产业啥啥啥是有关关关关关关关关系的
嗯/i/p三[三三三/r/b]四线城[城城/b城城/b]市吧	三三三三三四键长长长长长是吧
很多就是可能和以前是/p反[反反反反/r/b反]常识的一些观[观]点	很多就是可能和以前是反反反反反常识的一些观观点

Table 5. **Three examples of utterances from a severe PWS, characterized by frequent word repetitions. The wav2Vec model produced homophone substitutions, as indicated by the underlined characters.**

Annotation	Wav2Vec model output
当[当当当]时我上/b[上]去的时候	当当档单舍瓦上上去的时候
我/b现[现现现]就[就]挺自/r卑[卑]的	我先线先千就就点自杯给的
呃/i进[进]行那个自[自]我介[介]绍，呃/i	而仅仅进行了个自自我界介绍和儿
嗯/i/p我觉得非常好就嗯/i/p/r介绍完毕	我觉得黑套这结算完毕