Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

The Speech Accessibility Project: Best Practices for Collection and Curation of Disordered Speech

Chris Zwilling[1], *Mark Hasegawa-Johnson*[1], *Heather Hodges*[2], *Lorraine Ramig*[2], *Adina Bradshaw*[3], *Clarion Mendes*[1], *Heejin Kim*[1], *Alexandria Barkhimer*[2], *Laura Mattie*[1], *Meg Dickinson*[1], *Shawnise Carter*[4], *Marie Moore Channell*[1]

[1]University of Illinois, [2]Lee Silverman Voice Treatment, [3]Team Gleason, [4]Lingraphica,

zwillin1@illinois.edu, jhasegaw@illinois.edu, cmendes2@illinois.edu, hkim17@illinois.edu, barkhim2@illinois.edu, ljhahn@illinois.edu, megd@illinois.edu, channell@illinois.edu, heather.hodges@lsvtglobal.com, lori.ramig@gmail.com, Adina.Bradshaw@teamgleason.org, scarter@healthycarehomecare.com

Abstract

Drawing lessons from the collection and curation of disordered speech samples from a large open-source database that includes dysarthria, apraxia, dysphonia, or otherwise atypical speech—the Speech Accessibility Project—this paper provides best practices to guide the collection of diverse speech samples and enhance reproducibility. This paper describes data collection partnerships; the information technology architecture required to collect a massively distributed dataset spanning countries; the data collection process from the participant point of view; and the speech prompts and annotation process used for diverse speech. The open-source data set resulting from this set of best practices will provide a high-quality testbed for researchers around the world to train automatic speech recognition algorithms for disordered speech.

Index Terms: data collection, disordered speech, speech recognition, open-source database

1. Motivation

The Speech Accessibility Project corpus is a large collection of dysarthric, apraxic, dysphonic, or otherwise atypical speech recordings. It is designed to serve as a testbed for improving automated speech recognition (ASR) algorithms to equalize word error rates (WER) between dysarthric and non-dysarthric speech. WERs for non-dysarthric speech at the end of 2023 were less than 2% while WERs for dysarthric speech exceeded 18% [1]. Additionally, because individuals with dysarthric speech often have physical impairments, accurate ASRs are all the more valuable to promote accessibility to aid individuals in activities of daily living that rely on assistive speech technologies [2, 3].

2. A Network of Partnerships

The success of the Speech Accessibility Project, as measured by the number and quality of speech recordings collected, is due in large part to collaborative partnerships spanning seven distinct groups, involving more than 3,000 individuals, and 200 organizations. Figure 1 shows a network diagram of how some of these groups are interconnected and work together.

2.1. Core Team

Data collection was led by a core team of six professionals at the University of Illinois and consisted of the Principal Investigator, a Program Manager, an Information Technology coordinator, a Lead Mentor, a Lead Annotator, and Communications Director. The Core Team is the only group to directly interact with all other teams in the project.

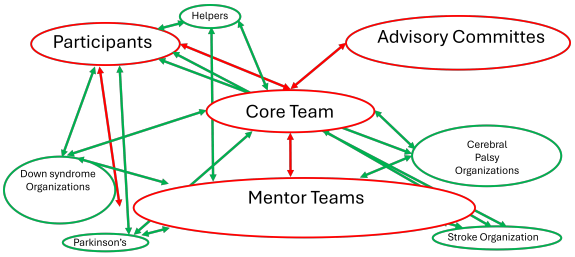


Figure 1: *Network diagram of Speech Accessibility partnerships.*

2.2. Mentor Teams

Methods for screening and mentoring participants were modelled after methods that have previously been demonstrated to be successful for telemonitoring of Parkinson's disease, including [4, 5]. Mentors, typically certified and licensed speech language pathologists, meet with potential participants via teleconference to determine if speech patterns meet criteria for the study. The Speech Accessibility Project partnered with five teams to lead mentoring, with each specializing in speech impairments due to one of the five etiologies included in the study. Lee Silverman Voice Treatment specializes in speech impairments due to Parkinson's Disease. Team Gleason works with individuals living with Amyotrophic Lateral Sclerosis, or ALS. Lingraphica has expertise with speech and language impairments due to stroke. And two teams at the University of Illinois, one with expertise in Down syndrome and the other with expertise in cerebral palsy, led mentoring for those etiologies. Altogether, the Speech Accessibility Project has more than 50 individuals who are part of the core mentor teams. The mentor teams interact directly with the Core Team, Recruitment Organizations, Participants and their Care Partners.

2.3. Recruitment Organizations

Each lead mentor team representing one of the five etiologies included in the Speech Accessbility Project connected with multiple organizations to promote recruitment. The lead mentor teams for both Down syndrome and Parkinson's Disease each partnered with more than 100 different organizations to promote

10.21437/Interspeech.2025-2454

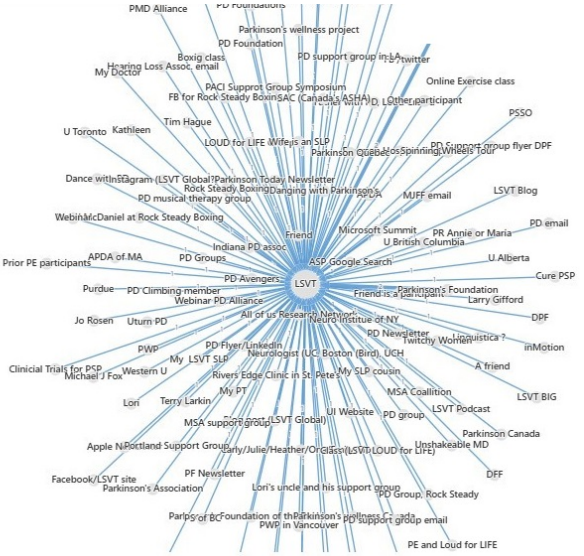


Figure 2: *Lee Silverman Voice Treatment recruitment network.*

the study through advertising, sharing of recruitment flyers, distributing communications through email list serves, and giving talks (see Figure 2 for Lee Silverman Voice Treatment's recruitment organizations). In these cases, each of their 100-plus recruitment partner organizations contacted typically brought a few participants into the study. By contrast, the cerebral palsy lead mentor team partnered with only five organizations; but each of those organizations had deep connections with participants, with each organization bringing approximately 75 participants into the study. Team Gleason has a strong connection with the ALS community with highly motivated members that want to participate in research and initiatives that support others facing ALS. Team Gleason connects with community members via 50+ touchpoints per week, sharing Speech Accessibility Project information during these interactions. ALS participants learned about Speech Accessibility Project through the Team Gleason mailing list, ALS clinics, social media, support groups, staff touch points, and peer word-of-mouth. The mentoring lead for stroke has an extensive mailing list of participants already connected to their organization and which served as their primary recruiting contacts.

2.3.1. Communications and Public Relations

The Communications Director designed flyers approved by the Institutional Review Board (IRB) to promote the project. Examples of two flyers are at https://tinyurl.com/3365zas5 and https://tinyurl.com/5pjr423h. These flyers are distributed to partner organizations and posted on their social media channels and sent through mailing lists to their recruitment partners. The Communications Director grew a mailing list to over 700 people to provide project updates about progress and upcoming milestones. The Communications Director and her communications colleagues from the Beckman Institute for Advanced Science and Technology also wrote stories, coordinated with mass media, assisted in setting up podcasts and interviews, worked directly with mentor organizations and their marketing teams, maintained the main project website (https://speechaccessibilityproject.beckman.illinois.edu/) and overall directly and indirectly supported the recruitment of study participants.

2.3.2. Recruitment Lesson Learned

The project experimented with a clinical recruiter to increase registrations. The recruiter screened leads delivered from postings on their social media channels, thereby serving as a buffer between the Speech Accessibility Project research team and a large amount of traffic from social media. Of the approximately 200 leads provided to the Speech Accessibility Project over three months, only two of those were eligible, enrolled and completed the study. All the other leads were dead ends, either because they didn't meet the screening requirements or because they were fabricating the symptoms of a speech impairment they did not possess. Fabrication of demographic attributes in order to falsify eligibility for a paid research study has been widely reported in other work, e.g., 45% of respondents in [6]. Recommended methods for detecting such fabrication include conversation with an expert in which the respondent answers questions that test the claimed symptom in ways not often understood by non-experts [7, 8]. Speech Accessibility Project mentor teams developed etiology-specific protocols to identify the fabrication of speech impairment, but there was always a loss of time and team morale in dealing with these false leads. On the whole, because of the large number of leads that did not convert to study enrollments, the team considers the work with the clinical recruiter a failed experiment. Working with known networks of individuals and organizations seems to be the best approach for recruiting participants with the etiologies included in the Speech Accessibility Project.

2.4. Participants and Care Partners

At the project's inception, there were three goals for the size and scope of the database: 1,000 hours of speech recordings (modeled after Librispeech [9]), 2,000 participants (double the diversity of Librispeech, because recent results confirm that participant diversity improves speaker-independent ASR, e.g., [10]), with 400 participants coming from each of the five etiologies. As of the writing of this paper, with 6 months remaining for data collection, there were 442 contributors with Parkinson's, 334 with ALS, 316 with cerebral palsy, 141 with Down syndrome, and 190 with stroke, for a total of more than 1,400 contributors since the project launched; the goal of 1,000 hours was passed in early 2025. More than half of those contributors had care partners, who assist the participants in different ways, depending on their needs, helping to read the prompts, assisting with the technology, and coordinating payment.

2.5. Annotation Team

The 20-plus person annotation team is responsible for transcribing all speech recordings. Individuals working on annotation have expertise in speech and hearing science, or a similar background.

2.6. Advisory Committees

The project has a core advisory committee and a technical advisory committee. The core advisory committee meets monthly, includes one representative from each of the funding partners, is responsible for setting the direction of the study, selecting etiologies to include, approving organizations to add as subcontractors, and making decisions to keep the project on track. The technical advisory committee is concerned with data analysis, providing feedback to the study team about the quality of data

and suggesting adjustments, providing guidance on matters of data usage for the project, and assisting with data challenges to promote the database. The technical advisory committee includes at least one member from each of the funding partners but often includes two or more members. The advisory and technical advisory committees include more than 20 members in total. The advisory committees only interact with the Core Team (see Figure 1).

2.7. Data Use Partners

Any individual who is part of an organization may request access to the Speech Accessibility Project data (https://tinyurl.com/yujrwxnv). The interested party fills out and signs a Data Use Agreement (DUA), which includes legal protections for participant data. The DUA is reviewed and signed by the University of Illinois Office of Technology Management, typically within three weeks of submission. The user who requested the data will then receive a link to the data along with instructions and guidance on how to use and access it. The Speech Accessibility Project has completed DUAs with more than 60 organizations and shared data with companies and academic partners from around the world, including users in the United States, India, Denmark, Japan, South Korea, Italy, Germany, New Zealand, United Kingdom, Russia, Canada, Vietnam and France.

3. IT Architecture

The Information Technology (IT) architecture consists of three interfaces: a contributor interface, an annotator interface, and an administrator interface. The data is stored in two databases: a Speech Database that is available to researchers and a Private Database only available to administrators. Each database is stored on a separate server for added security. Each database also has development, test, and production environments to permit prototyping changes before implementing them in the contributor interface. Data is also stored both on-site and off-site. File storage requirements are on the order of a few Terabytes for the datasets and metadata. A completed dataset for one participant includes 450 utterances, with between 1 and 180 seconds of audio per utterance.

The Speech Database has two tables: the contributor shareable metadata table and audio table. The metadata table includes metadata about each contributor that can be shared with any researcher who has signed a data use agreement. The contributor key is a unique identifier (UUID, [11]). Each contributor is associated with a small number of metadata fields, with each having "last modified" fields, specifying date and time in a standard format (e.g., ISO 8601). Metadata contributed by the participant includes only contributor identification of their own native language (if not English) and of the etiology of their speech impairment; other participant-contributed metadata is not included, to reduce the risk that a bad actor might gain some benefit from data aggregation [12]. Metadata contributed by the research team includes prompt text and transcription text for each audio file, and differential diagnostic patterns of dysarthria as annotated by speech language pathologist mentors for up to fifteen audio files per participant [13].

The Private Database for administrators is organized with a unique identifier (UUID) that ties it back to the individual in the Speech Database and metadata that includes the participant's state and/or country of residence, date of birth, gender, race and ethnicity if provided by the participant, email address, and cumulative number of contributions. Demographic information is preserved in the administrator table for the purpose of aggregate reporting; contact information is preserved so that human subject payments can be made appropriately.

Data contributors can easily request tech support by pressing a "Help" button that provides a template for trouble ticket submission. Tech support have a list of Frequently Asked Questions for which they can provide an answer and then route the remaining issues to the Program Management Lead.

4. Participant Access

Once an individual decides to to assess their eligibility for participating in the study, they register on the Speech Accessibility Project website (https://saa.beckman.illinois.edu/Identity/Account/DiagnosisRegister). During registration, the individual selects the etiology that gives rise to their speech impairment (Parkinson's Down syndrome, cerebral pasly, stroke or ALS), and provides some basic demographic information, including their state of residence. Individuals from three states—Illinois, Texas and Washington—in the United States are ineligible to participate due to biometric privacy laws (for the Illinois law, see here: https://tinyurl.com/wswwaywu).

A mentor then reviews the information collected during registration. If the individual is a potential fit for the study, a mentor contacts them to set up a virtual meeting to further assess speech profile, answer questions, and determine if the individual is eligible. If eligibility is met, the mentor helps the now participant, and if needed, the helper and guardian through the informed consent process (https://tinyurl.com/7ffwzr9s). All consent forms were approved by the University of Illinois IRB. The IRB also approved the entire study protocol and approves amendments to modify the study. If the participant has a helper, they are present during the initial mentor meeting. If at any point during the process the individual does not meet the criteria for participation (e.g., under 18 years of age or reside in an ineligible state), they are thanked for their interest but told they are ineligible.

Once a participant has completed the informed consent process, they may login to the contributor interface at any time of day, from any device, and from any location with internet connectivity. The main page of the contributor interface presents a text prompt, a menu bar and audio control buttons: "Record," "Stop," "Re-record," and "Save." In the normal flow of contribution, participants click the "Record" button, read the prompt sentence, then click "Stop." If they press "Save," then the recording will be saved to the database, and the existing prompt sentence will be replaced by a new prompt (see Figure 3).

Each participant is asked to read 350 - 400 sentences, and to respond to 50 - 80 questions for spontaneous speech with responses of approximately four sentences each, which are broken into 10 blocks. Each block includes prompts which the participant contributor reads verbatim (i.e., read speech) and a smaller number of spontaneous prompts, generally in the form of a question, where a participant responds however they like. When a contributor wishes to stop recording, they may simply close their browser. When they wish to begin recording again, the login URL will ask for their e-mail address and password, then take them immediately to the main contributor interface.

There is no set time frame or schedule under which participants must complete recordings; most complete within 3-5 recording sessions, and few fail to finish within 12 weeks. As long as participants make continued progress, login regu-

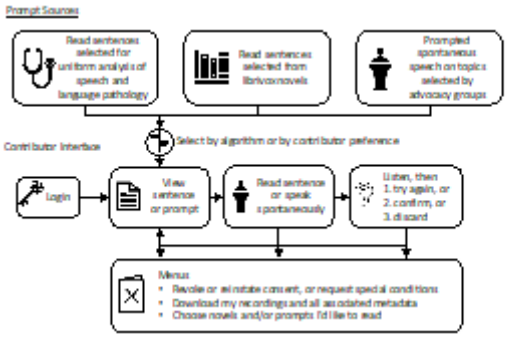


Figure 3: *Schematic block diagram of the contributor interface.*

larly, and communicate with their mentor, they can continue recording until all prompts are completed. If a participant stops recording and is non-responsive to mentor queries, their recordings are saved and included in the database but their participation is classified as discontinued. Participants are automatically paid $60 via Amazon e-codes after the 3rd, 7th, and 10th blocks for a maximum of $180. Helpers are eligible for $30 on the same payment schedule as participants, for a maximum of $90.

4.1. Meeting Diverse Participant Needs

The diverse etiologies presented challenges to which the project had to adapt to successfully meet the needs of the different communities. For example, we provided multiple ways for Down syndrome participants to access the prompts and participate, depending on an individual's support needs. Each participant could select from one of three options: (i) read the prompt on their own (ii) have a helper read it to them and then repeat it or (iii) listen to an automated voice reading the prompt via headphones and then repeat the prompt aloud. The latter option obviated the need for a human assistant to read each word while simultaneously preventing interference in the annotation because only the participant voice was heard, and not the reader. By contrast, individuals with Parkinson's Disease either read the prompts on their own or had a helper say the prompt and then repeat it.

5. Speech Prompts and Annotation

Participant recordings include both read and spontaneous prompts. The full list of prompts for each etiology is here: https://tinyurl.com/45w63tez. There are four prompt sets, which are assigned based on the etiology of the participant. All prompts were developed in conjunction with the Lead Mentor, Lead Annotator, the core Mentor Teams, and focus groups. The prompts for participants with Parkinson's and ALS are written at a 10th grade reading level and include named entities and concepts likely familiar to a person over fifty years of age. Down syndrome prompts are written at a 5th grade reading level and include named entities and concepts likely familiar to a person under fifty years of age. Most research participants with Down syndrome are under 50 years old and have intellectual disabilities so it is important the prompts align with their cognitive abilities. Stroke prompts were further modified from the set of Down syndrome prompts. Some participants with cerebral palsy, and in some cases, other diagnoses, may rely on augmentative and alternative communication and may have difficulty producing intelligible speech. In these cases, participants are offered the option of reading isolated words instead of responding to prompts for spontaneous speech. The distribution of prompts and subcategories included is listed within the GitHub repository link provided at the beginning of this section. Participants with ALS or stroke are also offered this option if their speech is severely impaired or they experience discomfort and profound fatigue with longer speech.

Each utterance is assessed by the Annotation Team. Transcriptions of read speech are initialized using prompt text, then modified as needed by the Annotation Team so that the transcription matches what the participant actually said. Spontaneous speech is transcribed verbatim. Personally Identifiable Information (e.g., their names, address) is removed to protect data privacy. In addition, empty recordings and those containing only voices other than the participant are removed from the database. Disfluent sections such as repetitions of words or syllables, false starts, and filler words are also marked, following the Transcription Guidelines, which can be viewed here: https://tinyurl.com/396d62e5. Finally, 20-30 utterances per participant are annotated using the SLP Annotation Descriptors, which can be viewed here: https://tinyurl.com/yc8a97j8. Certified speech-language pathologists further classify two types of speech pathology annotations. First, the speech sample will be annotated with dysarthria type (flaccid, spastic, ataxic, hypokinetic, hyperkinetic, or a mixture of any two or more of these, other speech disorder, or no speech disorder). Second, the speech sample will be annotated using the Darley, Aronson, and Brown (DAB, [13]) differential diagnostic pattern scales. DAB is a set of 46 perceptual attribute scales, including scales related to articulation, intonation, loudness, voice quality, and overall intelligibility. Every contributed utterance, read speech or responses to spontaneous prompts, is checked by a member of the Annotation Team.

6. Discussion

The Speech Accessibility Project corpus is an important testbed for disordered speech and aims to enable automated speech recognition technologies to achieve word error rates comparable to non-disordered speech. The scope, scale, and diversity of the speech recordings in the database, which will include speech contributions from 2,000 participants with speaking impairments due to five different etiologies, has the potential to make this goal a reality. Moreover, the quality of the speech recordings is ensured because of the processes and methods implemented to collect the speech samples, all of which have been outlined in this paper. The interpersonal and organizational collaborations, IT architecture, participant interface, and annotation procedures all contribute to the quality of this database and provide a set of best practices for other researchers to follow when designing protocols to collect speech recordings.

7. Acknowledgements

We thank the AI Accessibility Coalition comprised of Amazon, Apple, Google, Meta and Microsoft.

8. References

[1] M. Hasegawa-Johnson, X. Zheng, H. Kim, C. Mendes, M. Dickinson, E. Hege, C. Zwilling, M. M. Channell, L. Mattie, H. Hodges *et al.*, "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4162–4175, 2024.

[2] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people," *Arch. Phys. Med. Rehabil.*, vol. 66, pp. 678–681, 1985.

[3] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria." *Med Eng Phys*, vol. 29, no. 5, pp. 586–593, 2007.

[4] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests." *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[5] R. Fraile, N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, "Mfcc-based remote pathology detection on speech transmitted through the telephone channel-impact of linear distortions: Band limitation, frequency response and noise." *International Conference on Bio-inspired Systems and Signal Processing*, vol. 1, pp. 41–48, 2009.

[6] J. J. Chandler and G. Paolacci, "Lie for a dime: When most prescreening responses are honest but most study participants are impostors." *Social Psychological and Personality Science*, vol. 8, no. 5, pp. 500–508, 2017.

[7] I. P. Kan and A. B. Drummey, "Do imposters threaten data quality? an examination of worker misrepresentation and downstream consequences in amazon's mechanical turk workforce." *Computers in Human Behavior*, vol. 83, pp. 243–253, 2018.

[8] D. N. Shapiro, J. Chandler, , and P. A. Mueller, "Using mechanical turk to study clinical populations." *Clinical Psychological Science*, vol. 1, no. 2, pp. 213–220, 2013.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books." *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, 2015.

[10] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, "Training and typological bias in asr performance for world englishes." *Proc. Interspeech 2022*, pp. 1273–1277, 2022.

[11] P. Leach, M. Mealling, and R. Salz, "Rfc 4122: A universally unique identifier (uuid) urn namespace." Microsoft, Tech. Rep., 2005.

[12] G. D. . N. Zhang, "Privacy risks in health databases from aggregate disclosure." *Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 1–4, 2009.

[13] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria." *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246–269, 1969.