Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing

A multi-project perspective

Yue, Zhengjun; Barberis, Mara; Patel, Tanvina; Dineley, Judith; Doedens, Willemijn; Stipdonk, Lottie; Zhang, Yuanyuan; De Witte, Elke; Scharenborg, Odette; More Authors

DOI
10.21437/Interspeech.2025-2774

Publication date
2025

Document Version
Final published version

Published in
Proc. Interspeech 2025

Citation (APA)
Yue, Z., Barberis, M., Patel, T., Dineley, J., Doedens, W., Stipdonk, L., Zhang, Y., De Witte, E., Scharenborg, O., & More Authors (2025). Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing: A multi-project perspective. In *Proc. Interspeech 2025* (pp. 3943-3947). (Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2025-2774

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

Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing: A multi-project perspective

Zhengjun Yue[1], Mara Barberis[2], Tanvina Patel[1,3], Judith Dineley[4], Willemijn Doedens[5], Lottie Stipdonk[3], Yuanyuan Zhang[1], Elke De Witte[3], Erfan Loweimi[4], Hugo Van hamme[2], Djaina Satoer[3], Marina Ruiter[6], Laureano Moro Velazquez[7], Nicholas Cummins[4], Odette Scharenborg[1]

[1]TU Delft, Netherlands; [2]KU Leuven, Belgium; [3]Erasmus MC, Netherlands; [4]KCL, UK; [5]Royal Dutch Auris Group, Netherlands; [6]Radboud University, Netherlands; [7]John Hopkins, USA

z.yue@tudelft.nl

Abstract

Speech technologies have advanced significantly, yet they remain largely trained on typical speech, limiting their applicability to individuals with speech and language impairments. A key obstacle is the lack of well-annotated and representative atypical speech corpora. This paper conducts a multi-project survey and shares the first-hand experience on the challenges of collecting, annotating, using, and sharing atypical speech data. Experiences from seven research projects on collecting atypical speech data, involving both academic and clinical perspectives, are reported and potential issues are discussed. Furthermore, the paper provides practical guidelines that allow for standardisation and harmonisation of data collection practices, which are crucial to allow studies to be compared, replicated, and validated, which is essential for developing more inclusive and effective speech technologies.

Index Terms: Dutch atypical speech, speech data collection, speech annotation, automatic speech recognition

1. Introduction

Speech technologies have advanced greatly in recent years due to, amongst others, the availability of increasingly larger data sources. However, the underlying models are, almost universally, trained on typical speech [1, 2, 3]. Such models tend not to generalize well to atypical speech produced by individuals with language and/or speech disorders [4, 5]. A major factor behind this lack of generalizability is insufficient well-annotated atypical speech corpora [6, 7]. Indeed, data and data collection form the foundation of almost all speech research [8] because without data research would be speculative rather than evidence-based. Despite this, there is no accepted standard for collecting or reporting on databases [9]. Reporting on data collection is a severely neglected aspect of speech processing research. More surprisingly, given the overwhelming need, it is difficult to obtain funding for speech data collection. This is particularly difficult in speech processing, which like many other research fields, faces a reproducibility crisis [10, 11]. This lack of transparency on data collection and use can lead to issues with reproducibility, data bias, and methodological inconsistencies.

Data collection challenges are broad, encompassing issues including recruitment, annotation, storage, and sharing. For example, many research projects are initiated and conducted within academic settings. It can be considerably more challenging to recruit speakers with disordered speech from an academic setting than from a medical setting [12]. Due to the difference in proximity and lack of direct access of the clients/patients, academic institutions often find it difficult to identify and engage potential participants [13]. As a result, data collection becomes more time-consuming, costly, and relatively vulnerable [9, 14]. Once collected, speech data must be accurately annotated, a process that is time-consuming, expensive, and highly subjective, particularly when dealing with disordered speech [15, 16]. Annotation consistency and validation remain critical challenges, as different annotators may interpret speech characteristics differently. As automated annotation tools [17, 18, 19] and pre-trained ASR models perform poorly on atypical speech, manual intervention is necessary. After datasets are created, sharing and reusing the data are restricted due to GDPR regulations, IRB limitations, voice anonymization challenges [20], and a lack of standardized sharing frameworks.

Perhaps the even bigger issue is "*disciplinary siloing*" between clinicians and engineers and between institutes, i.e., everyone creating their own processes according to their own research training and goals [21]. In a field where sharing data poses challenges due to speech's inherent identifiability, harmonisation and standardisation of processes on collecting, annotating, using, and reporting on speech data is vital to allow studies to be compared, replicated, and validated without the need to share data. Such a step would allow for meta-analysis and stronger scientific evidence.

Herein, we discuss the challenges associated with collecting, annotating, using, and sharing atypical speech databases. We aim to make recommendations about collecting and reporting on databases based on the authors' combined experiences in multiple, separate projects [22]. The common theme across these projects is that we have applied speech technologies, particularly ASR, to atypical speech. These projects on congenital or acquired language and/or speech disorders range from working with dysarthria [23], aphasia [24], children with developmental language disorder (DLD) [25], to stuttering, and are often collaborations between academics and speech therapists, hospitals, and clinical researchers. We present a thematic review of lessons learnt from seven projects that involved collecting atypical speech. Reviewing past challenges is a vital step in working towards effective and community-accepted solutions and will enhance the reliability and/or validity of data collected in future projects by being able to build on our lessons. The second aim is to - based on the lessons learnt - share practical guidelines that would allow for standardisation and harmonisation, which is crucial to allow studies to be compared, replicated, and validated. The findings presented here will be relevant to researchers working in language and speech technology, linguistics, and clinical applications, as well as those involved in ASR development for underrepresented speech populations.

2. Insights from research projects

This section draws insights from seven research projects that involve collecting, annotating and using atypical speech data

10.21437/Interspeech.2025-2774

Table 1: *Overview of the seven atypical speech research projects.*

Project# Name	Age (years)	Lang. (s)	Speech Type	Atypical Speech	#Speakers
P1 (Oral Speech Motor Skills Assessment)	children: 3;5-10;0 years	Dutch	Repetitive non-words	Stuttering	110
P2 (Development Language Disorder (DLD))	children: 1;0-12;0 years	Dutch	Read, spontaneous	Autistic, Down Syndrome, DLD	∼300
P3 (CHILDES for ASR)	children: 4;0-7;0 years	Dutch	Spontaneous speech	DLD	Ongoing
P4 (Aphasia Speech Recognition and Assessment)	Adults	Dutch	Natural speech	Aphasia	120
P5 (Awake Brain Surgery)	Adults	Dutch	Spontaneous speech	Pre/during/post Brain tumor	108
P6 (ASR for Dysarthria)	Adults	Dutch-Eng	Multi-modal (A+V)	Dysarthria	1
P7 (Speech Language Understanding (SLU)	Adults	English	Conversational	Neurological Disorders: PD, ALS, MS	Ongoing

across different populations, including children and adults with speech and language disorders. A brief overview of the various projects and their details is provided in Table 1.

Project 1 aims to assess oral speech motor skills in children who stutter and fluent controls by recording repetitive non-word sequences. Project 2 is using the existing CHILDES [26] multilingual corpus, which consists of various types of child speech, widely used for linguistic research for ASR training. Projects 3, 4, and 5 have similar aims to 1) automate speech recognition and/or assessment of (semi-) spontaneous speech data from children with DLD (P3), of natural speech from individuals with aphasia (P4), and of spontaneous speech from patients undergoing awake brain tumor surgery collected before, during, and after the surgery; and potentially 2) explore other patterns/markers in speech for clinical purposes. Project 6 focuses on developing personalized ASR systems for people with dysarthria, and collected multi-modal (audio and video) and Dutch-English bilingual data from a Dutch speaker. Project 7 is aimed at improving spoken language understanding (SLU) systems for people with atypical speech in English, and data collection includes adults with various neurological disorders. In this section, instead of listing challenges broadly, we identify unique and/or shared key issues/themes in projects, emphasizing what was learned from addressing these difficulties.

2.1. P1/P3/P4/P5/P7: Clinical Speech Data Collection

Speech data collection is a fundamental step in developing atypical speech datasets; however it comes with several challenges. This section discusses P1, P3, P4, P5, and P7 which share common challenges and insights although they differ in their target populations and data collection setups.

The first challenge is concerned with ensuring participant engagement and eliciting sufficient speech for each speaker. In order to ensure that a sufficient amount of speech is elicited from each speaker, the therapist often needs to coax and train the speaker, which is also recorded. This problem particularly arose in those projects that are concerned with child speech and adults with stroke-induced aphasia. The speech of the therapist needs to be removed from the participant's speech files. Moreover, the therapist's speech might overlap with the participant's speech, which would reduce the amount of data that can be used for further processing. Unnecessary interference from the test administrator might sometimes disrupt the flow of natural discourse, leading to speech patterns that are less representative of real-life communication. Conversely, a lack of interference from the test administrator while leading to "cleaner" data, can also lead to less natural conversation. Speakers might also interrupt their own speech to correct themselves or because they are distracted. P1 faced this issue with their child speakers who frequently interrupted their own speech sequences, were more prone to external distractions, and were inconsistent in their task execution, making it difficult to obtain reliable, uninterrupted samples needed for motor skill assessment. Neurologically impaired adults and adults with brain tumors often have comorbid cognitive deficits (i.e., the co-occurrence of two or more speech and/or language disorders) as in P4 and P5, which results in similar problems as those in the child speech projects, and can further complicate automatic recognition of atypical speech. Another major challenge is physical movement and body positioning during recordings. Children, for instance, tend to move excessively, place their hands in front of their mouths, or lean away from the microphone, which influences audio clarity and quality. This can also occur in individuals with postural instability or tremors, common in some neurological disorders. Such movement-induced noise introduced acoustic artifacts, making it difficult for both human transcribers and ASR models to reliably transcribe the speech.

Furthermore, the recording environment has an impact on speech data collection, particularly in clinical settings. For instance, recordings made in hospital environments can be affected by background noise from medical equipment (e.g., machines bleeping) and medical conditions (e.g., excessive breathing sounds (wheezing)) impacting the quality and usability of the collected speech. In P1, for instance, unintentional noise from both the child participants and researchers made automatic sequence analysis less reliable, requiring manual intervention to ensure accuracy. Similar problems arise when a data collection team lacks the expertise to standardize data collection settings across different recording sessions, yielding recordings with inconsistent quality.

A main takeaway from these challenges faced by several projects is that participation incentives matter. Proper speech elicitation is needed during recording, and it needs to be tailored to accommodate the specific challenges of each participant group. For instance, researchers from P4 suggest that compared with fully open-ended prompts, incorporating semi-structured discourse tasks such as picture descriptions, procedural explanations, or narrating silent films can encourage more structured yet natural speech production. These structured but flexible approaches can help to at least partially mitigate the common issue of therapist or researcher unnecessary intervention in speech recordings while still allowing for natural variation in speech output. For P1, for example, we recommend that child-friendly game-based prompts or interactive activities can help maintain children's engagement as well as attention, and reduce spontaneous interruptions during speech production. Clear pre-recording instructions and practice sessions reinforce the need for continuous speech production.

Additionally, well-controlled recording environments and setups with minimal distractions and using high-quality microphones can help reduce ambient noise. Especially in medical settings learnt from P4 and P5, using directional microphones and adaptive noise filtering techniques can help mitigate background noise. Appropriate noise-reduction methods can also further enhance speech data quality in the pre-processing stage.

Headset or lapel microphones, used in P7 [27], are useful to maintain mouth-to-microphone distance constant and similar to all participants, which has benefits for other types of studies about atypical speech quality measurements.

It is worth mentioning that it is important to maintain realism in speech recordings. For example, the unintended movements that interfere with audio clarity, while minimizing movement may improve recording quality, overly controlled conditions could lead to unrealistic speech data that does not reflect natural behavior. This limitation is also relevant for headset microphones, which provide cleaner recordings but fail to capture the variability introduced by natural body movements. That is, one needs to always ensure that the training data is collected in a similar situation as the eventual system will be used. Creating "clean" data that is different from the setting in which the system will be deployed is counter-productive. Therefore, a balance must be struck between optimizing recording quality and maintaining ecological validity to ensure that ASR models generalize well to real-world speech conditions.

2.2. P3/P4/P6: Natural/spontaneous speech annotation

A major challenge is the annotation of natural and (semi-)spontaneous speech. Unlike scripted speech tasks, conversational and spontaneous speech lack reference prompts, and are highly unpredictable and variable, characterized by frequent pauses, hesitations, non-words, mispronunciations, and self-corrections, requiring extra/further time-consuming and labor-resource-intensive manual transcription for those variabilities. Moreover, different needs for the transcriptions for speech technology and language research makes it challenging to create a standardized annotation protocol. Atypical speech increases complexity by featuring irregular prosody, fragmented syntax, and ambiguous articulation. In P6, for example, the dysarthric speaker's speech distortions and fatigue-related variations required multiple rounds of listening and correction before an accurate transcription could be finalized. Similarly, in P3 and P4, children with DLD and individuals with aphasia as well as comorbid speech disorders frequently produced non-standard utterances that make traditional annotation methods difficult.

Moreover, it was noticed that inconsistencies in transcription happened due to annotators having different interpretations of the same sound or utterance. For instance, in P6, subtle phonetic variations or barely audible sounds, such as /p/, /s/, /t/, or /k/ at the end of some isolated words were sometimes omitted, requiring later revisions. In P4, missing words or unintelligible segments had to be carefully and manually marked to ensure that aphasic speech characteristics were accurately captured. Additionally, since atypical speech is inherently more ambiguous than typical speech, transcribing spontaneous speech requires specialized linguistic expertise, which does not ensure the correct annotation of highly disordered speech. Multiple annotation passes increase the time and cost.

A specific challenge in P3 and P4 was the multiple interactions between the test administrator and the participant. These interactions, which included prompting, clarifications, and conversational support, needed to be carefully annotated to differentiate speaker turns and capture the natural discourse, i.e., "who speaks what and when". This separation will be particularly useful when training ASR systems using only atypical speech parts in the dataset.

A key takeaway from these challenges is the importance of iterative annotation workflows, similar to inter-rater reliability measures used for transcription in language research. In P6,

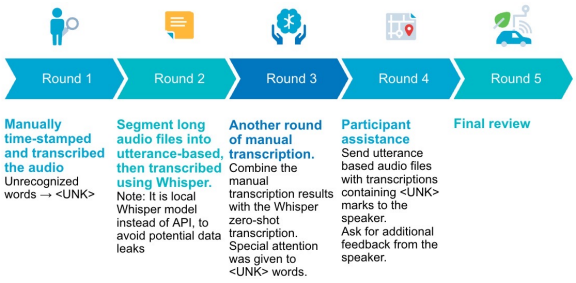

Figure 1: *The process of spontaneous speech annotation in P6.*

an approach involving multiple rounds of transcription and participant' verification led to high confidence of completing the dataset, as shown in Fig. 1. Semi-automated annotation tools played an important role in reducing manual effort. In P4 and P6, ASR-assisted transcription (e.g., Whisper [28]) provided an initial or second draft, which was then refined by human annotators. While ASR performance on atypical speech remains suboptimal, leveraging pre-trained models for preliminary transcriptions helped accelerate the annotation process (e.g., 4 times faster compared to a manual approach reported in [29]) by reducing the amount of text that is needed to be transcribed manually from scratch, which improves both efficiency and reliability. Moreover, pre-defined guidelines for handling detailed information such as pauses, hesitations, and unintelligible speech can ensure the consistency of the transcription across annotators. The team from P4 suggested starting from an existing protocol (e.g., Protocol voor Orthografische Transcriptie [30]) and adapting it for the type of atypical speech [31].

Another crucial lesson is the value of speaker-assisted annotation/feedback. The dysarthric speaker in P6. was consulted to clarify unintelligible segments marked by annotators, helping annotators distinguish between mispronunciations, phonetic variations, and word substitutions. This approach significantly improved transcription accuracy and ensured that the dataset more faithfully represented the speaker's speech. However, this procedure takes participants' time and effort and it might not be guaranteed to be possible with all participants (and most probably only a few of them would agree to do so). Alternatively, we can consider designing the type of speech to be recorded, e.g., semi-structured discourse tasks can help balance naturalness with annotation feasibility (but see the previous section). Structured yet open-ended speech samples that retain spontaneity while making transcription and analysis more manageabl.

For hard(er)-to-understand speech, it might be efficient to have the speech transcribed by someone familiar with the type of speech (e.g., the dysarthric speech in P6) or the specific speaker, as familiarity with the speaker and speech helps recognition over time [32]. Moreover, in the early transcription stages, involving multiple annotators can provide diverse perspectives on difficult-to-understand segments before finalising the transcription by a single annotator. Annotation quality control is also essential, as the absence of a standardized evaluation pipeline initially led to transcription errors in P6, such as the omission of barely audible sounds. To address this, pilot ASR experiments were used to identify errors, prompting a review of transcriptions. Additionally, misspellings can affect both linguistic analysis and ASR performance; therefore, lexicons from standard datasets (e.g., CGN [33], LibriSpeech [34]) were used to detect and correct errors, with further verification from the speaker when necessary.

2.3. P2/P4: Adaptation of linguistic datasets for ASR

The Child Language Data Exchange System (CHILDES) [26] is a well-established corpus designed for linguistic research and language acquisition studies. It offers annotated transcripts of different types of child speech in multiple languages, making it valuable for developing child ASR systems. However, CHILDES was not originally intended for ASR training, and repurposing it for speech recognition applications requires extra preprocessing and adaptation to improve its usability. The CHILDES corpus follows a structured linguistic annotation system called Codes for the Human Analysis of Transcripts (CHAT) [35], originally designed for language acquisition studies rather than automatic speech processing. While this structured format is beneficial for linguistic analysis, it introduces several challenges when applying the dataset for ASR training. In the datasets originally designed for ASR [34, 36, 37], each speech sample is typically paired with its transcription. However, in CHILDES, audio recordings and transcripts are stored separately. This requires additional effort to organize, align, and properly synchronize them before use.

Additionally, the transcription style used in CHILDES differs from ASR-ready datasets. It includes elements such as:

- Speaker labels (e.g., "CHI" for child, "INV" for investigator), which are not necessary for ASR.
- Grammatical tagging and phonetic variations (e.g., morphosyntactic tags and syntactic corrections), which must be removed before ASR training.
- Repetitions, disfluencies, and non-verbal markers (e.g., pauses, hesitations, and prosodic cues), which can interfere with ASR models if not properly handled.

Detailed annotations are valuable for linguistic research but introduce inconsistencies in formatting that must be removed or reformatted before using it for ASR, which was also experienced in P4's annotation protocol [31]. Without proper handling, it can introduce training biases, leading to poor ASR performance. Apart from that, annotation variability across annotators is also a great challenge. Some transcribers focus on stuttering or speech sound disorders, while others annotate dialectical variations. This diversity results in inconsistent transcription formats which makes it challenging for ASR training [38].

To effectively use existing datasets, such as CHILDES, for ASR, often means that structured preprocessing is necessary to ensure that ASR models learn meaningful speech patterns rather than being affected by annotation inconsistencies. Customized preprocessing can filter out dataset-specific annotations. Furthermore, running preliminary ASR experiments on preprocessed CHILDES data can help identify transcription errors and misalignment issues, similar to P6's approach, allowing researchers to validate and refine the dataset before being used into large-scale training. Additionally, data collection needs to be adapted, or a generalized pipeline is needed to be able to utilize such data for ASR modelling despite the inconsistencies.

3. Multidisciplinary Perspectives

To gain multi-disciplinary insights into data collection challenges, we distributed a questionnaire to fifteen researchers from different disciplines (engineers, computer scientists, speech and language therapists, and clinicians) representing seven projects, which included some form of atypical speech collection. The questionnaire asked respondents about (i) specific data collection challenges encountered in a recent project, (ii) the possible impacts of these challenges, and (iii) relevant suggestions and workarounds to address the issues. Figure 2 presents a word cloud of the questionnaire responses. The most frequently given concerns were ethical and legal issues (7 projects), issues with participant recruitment (6 projects) and annotation challenges (6 projects).

Figure 2: *Word cloud for response of questionnaire.*

Ethical and legal concerns highlight the challenges of GDPR and institutional restrictions. While these regulations are vital for safeguarding patient rights and data security, they can be perceived as limiting data availability, collection, usage, and distribution. Researchers must navigate complex approvals and develop ethical data-sharing frameworks to balance privacy with scientific progress.

Participant recruitment challenges were a common theme across the projects. Recruitment of speakers with impaired speech and/or language is often complicated by the sensitive nature of their condition. Improved participant experiences through patient and public involvement and engagement (PPI/E) [39] was given as a specific example to address this concern. There is an urgent need to work with research participants to understand their concerns with participation, e. g. comfort with different speech tasks [40]; and sharing data. PPI/E will help develop data collection processes that go beyond ethical/legal requirements and have good public buy-in.

Annotating atypical speech presents multiple challenges, including high costs, time demands, and subjectivity [41]. Questionnaire responses highlighted gaps in automated annotation workflows (5 projects) and ASR adaptation (5 projects). A clear need exists for robust ASR models for automated transcription of atypical speech. At the same time, concerns about data sufficiency (4 projects) show a need for guidelines on the minimum dataset size required to train effective ASR models. High inter- and intra-speaker variability complicates this further, especially in small atypical speech corpora, showing the necessity for large datasets of atypical speech. Limited data may fail to capture the full range of variability, resulting in poor model generalization.

4. Conclusion

This paper highlights the clear need for more atypical speech samples in speech-processing research. However, it is also clear these data should be collected using agreed-upon standards, guided by participant input. Ultimately, greater transparency and consensus on data collection practices will benefit all speech researchers and facilitate the development of more accessible technologies.

5. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech*, 2021, pp. 2426–2430.

[3] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Interspeech*, 2012, pp. 1776–1779.

[5] S. Leivaditi, T. Matsushima, M. Coler, S. Nayak, and V. Verkhodanova, "Fine-tuning strategies for dutch dysarthric speech recognition: Evaluating the impact of healthy, disease-specific, and speaker-specific data," in *Interspeech*, 2024, pp. 1295–1299.

[6] T. Patel and O. Scharenborg, "Using data augmentations and vtln to reduce bias in dutch end-to-end speech recognition systems," *arXiv preprint arXiv:2307.02009*, 2023.

[7] Y. Zhang, Z. Yue, T. Patel, and O. Scharenborg, "Improving child speech recognition with augmented child-like speech," in *Interspeech*, 2024, pp. 5183–5187.

[8] J. C. Félix-Brasdefer, "Data collection methods in speech act performance," *Speech act performance: Theoretical, empirical and methodological issues*, vol. 26, p. 41, 2010.

[9] O. Niebuhr and A. Michaud, "Speech data acquisition: the underestimated challenge," *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, vol. 3, pp. 1–42, 2015.

[10] M. Baker, "Reproducibility crisis," *nature*, 2016.

[11] T. Miyakawa, "No raw data, no science: another possible source of the reproducibility crisis," pp. 1–6, 2020.

[12] C. Fougeron *et al.*, "Developing an acoustic-phonetic characterization of dysarthric speech in french," in *7th International Conference on Language Resources, Technologies and Evaluation (LREC)*, vol. 1, no. 1, 2010, pp. 2831–2838.

[13] M. M. Baese-Berk, T. Bent, and E. Ryherd, "Communication in Medical Settings," 2024. [Online]. Available: https://acousticstoday.org/wp-content/uploads/2024/07/AT3-Communicating_featured_summer2024.pdf

[14] C. Cieri, D. Miller, and K. Walker, "Research methodologies, observations and outcomes in (conversational) speech data collection," in *Proc. HLT*, 2002.

[15] M. Corrales-Astorgano, D. Escudero-Mancebo, L. Aguilar, V. Flores-Lucas, V. Cardeñoso-Payo, C. Vivaracho-Pascual, and C. González-Ferreras, "A comprehensive rubric for annotating pathological speech," *arXiv preprint arXiv:2404.18851*, 2024.

[16] L. Armstrong, M. Brady, C. Mackenzie, and J. Norrie, "Transcription-less analysis of aphasic discourse: A clinician's dream or a possibility?" *Aphasiology*, 2007.

[17] SuperAnnotate, "SuperAnnotate - AI-Powered Audio Annotation Tool," 2025, [Accessed: 18-Feb-2025]. [Online]. Available: https://www.superannotate.com/audio-annotation

[18] Midas Research, "Audino - Open Source Audio Annotation Tool," 2025. [Online]. Available: https://github.com/midas-research/audino

[19] SPPAS Project, "SPPAS - Automatic Annotation of Speech Corpora," 2025. [Online]. Available: https://sppas.org

[20] A. Moretón and A. Jaramillo, "Anonymisation and re-identification risk for voice data," *Eur. Data Prot. L. Rev.*, 2021.

[21] N. Cummins *et al.*, "A methodological framework and exemplar protocol for the collection and analysis of repeated speech samples," *under review JMIR Research Protocols*, 2025.

[22] N. Cummins, L. L. White, Z. Rahman, C. Lucas, T. Pan, E. Carr, F. Matcham, J. Downs, R. J. Dobson, T. F. Quatieri *et al.*, "A methodological framework and exemplar protocol for the collection and analysis of repeated speech samples," 2024.

[23] 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.

[24] M. T. Sarno, M. Silverman, and E. Sands, "Speech therapy and language recovery in severe aphasia," *Journal of Speech and Hearing Research*, vol. 13, no. 3, pp. 607–623, 1970.

[25] D. V. M. Bishop and A. Edmundson, "Language-impaired 4-year-olds: Distinguishing transient from persistent impairment," *Journal of speech and hearing disorders*, vol. 52, 1987.

[26] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[27] H. Wang, V. Ravichandran, M. Rao, B. Lammers, M. Sydnor, N. Maragakis, A. A. Butala, J. Zhang, L. Clawson, V. Chovaz *et al.*, "Improving fairness for spoken language understanding in atypical speech with text-to-speech," in *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[29] M. Barberis, P. De Clercq, B. Tamm, H. Van hamme, and M. Vandermosten, "Automatic recognition and detection of aphasic natural speech," *Interspeech Proceedings*, pp. 1990–1994, 2024.

[30] N. Oostdijk, "Het corpus gesproken nederlands," *Nederlandse Taalkunde*, vol. 5, no. 3, pp. 280–284, 2000.

[31] Barberis, Mara, Van hamme, Hugo and Vandermosten, Maaike, "Protocol transcriptie van natuurlijke spraak bij afasie," 2023. [Online]. Available: https://osf.io/pmgwc/?view_only=77dddda4789d4343ab0b1e8b4cf4e0a7

[32] P. Drozdova, R. Van Hout, and O. Scharenborg, "L2 voice recognition: The role of speaker-, listener-, and stimulus-related factors," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3058–3068, 2017.

[33] I. Schuurman, M. Schouppe, H. Hoekstra, and T. Van der Wouden, "Cgn, an annotated corpus of spoken dutch," in *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL*, 2003.

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

[35] C. Pye, "The childes project: tools for analyzing talk," 1994.

[36] F. Rudzicz, A. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[37] C. Cucchiarini, H. V. Hamme, O. v. Herwijnen, and F. Smits, "Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," 2006.

[38] R. Sharma, D. Liu, J. Sun, S. Zhou, J. Qin, J. Xiong, and C. Chen, "Kidspeak: A general multi-purpose llm for kids' speech recognition and screening," *ICLR Conference*, 2025.

[39] R. Baines, H. Bradwell, K. Edwards, S. Stevens, S. Prime, J. Tredinnick-Rowe, M. Sibley, and A. Chatterjee, "Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review," *Health Expectations*, vol. 25, no. 4, pp. 1232–1245, 2022.

[40] J. Dineley *et al.*, "Remote smartphone-based speech collection: Acceptance and barriers in individuals with major depressive disorder," in *Interspeech*, 2Brno, Czechia,, 2021, pp. 631–635.

[41] M. Nicolao, H. Christensen, S. Cunningham, P. Green, and T. Hain, "A framework for collecting realistic recordings of dysarthric speech - the homeService corpus," in *LREC)*, May 2016.