DOI: 10.61782/fa.2025.0168

FORUM ACUSTICUM
EURONOISE 2025

AUTOMATIC SPEECH RECOGNITION FOR A DYSARTHRIC CHILD SPEAKING AUSTRIAN GERMAN

Lucas Eckert[1] **Barbara Schuppler** [1*]

[1] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

ABSTRACT

This paper compares Automatic Speech Recognition (ASR) systems for dysarthric child speech in Austrian German, focusing on a case of ataxic dysarthria. While dysarthria is well studied in adults, research on children is limited and no speech databases exist for dysarthric child speech in German, posing unique challenges for ASR development. In collaboration with the child, the family and the child's speech therapist, we decided how to record and annotate speech material of different styles, including read stories, digits, calculations and spontaneous dialogues. Using this material, experiments were conducted with different state-of-the-art ASR models, such as Whisper and Wav2Vec, applying finetuning and speech augmentation to address the limited dataset. Additionally, a recording tool was developed so the child can record new material in a familiar environment. Given that the ASR system shall be integrated into a real-time assistive technology, the next step will involve evaluating the ASR systems in real-life scenarios with the child to determine the most suitable option for daily use. This work demonstrates how data augmentation, tailored system adjustments, and collaborative approaches can address resource-constrained scenarios. The findings contribute to developing more inclusive ASR technologies for children with speech impairments.

Keywords: *dysarthric child speech, automatic speech recognition, assistive technology*

1. INTRODUCTION

Automatic Speech Recognition (ASR) has made significant advancements in recent years, achieving high accuracy for typical speech thanks to deep learning models like Whisper [1] and Wav2Vec 2.0 [2]. These systems perform well in acoustically diverse environments and can be fine-tuned to include speakers across different speaking styles — as long as sufficient training data is available. However, performance drops sharply when training data is limited, particularly for speakers with atypical speech patterns due to regional dialects or speaking impairments. This paper presents different ASR systems for dysarthric speech from a child speaking Austrian German.

When ASR is part of an assistive technology, it is not enough to train and evaluate models with global performance measures such as word error rate (WER). During all stages of tool creation, we need to consider the social environment of the dysarthric child and its developmental status. First and foremost, a speech disorder causes considerable restrictions in the child's everyday life. Making contact with other people is a major challenge for many when communication is characterized by misunderstandings, repetitions or the inability to communicate concerns [3]. Assistive communication technologies can support social interaction by overcoming language barriers, but they must be designed in a way that respects the individual's needs and avoids introducing new obstacles or enabling unnatural communication patterns that hinder interaction with peers [4]. Second, when collaborating with the relatives of children with communication disorders, it must be borne in mind that data acquisition must be conducted in such a way that it is neither too time-consuming nor requires too much (physical) effort, ideally as part of every-day activities. Third, engineers and speech thera-

[*] ***Corresponding author***: *b.schuppler@tugraz.at.*

pists need to collaborate to make design decisions such as whether the aid should synthesized complete utterances which "replace" the child's speech, or whether the technical aid should only support communication (e.g., by providing a transcript of the spoken words to increase intelligibility). This is the case that we deal with here, as given the speech therapists assessment, the child's pronunciation and language development could suffer from less every-day speaking practise which in turn could create an even stronger dependency on the assistive tool.

Concretely, this paper describes the development of an ASR system tailored to one specific Austrian German ataxic-dysarthric child speech. Thereby, two objectives arise. The first is to present our data collection process designed together with the family members and the speech therapist of the child, which allows to gather child speech in such a way that the process is also manageable for the subject. The second objective is to fine-tune state-of-the-art ASR systems, comparing the impact of different parameters and speech augmentation techniques on recognition performance. The long-term goal is that the ASR will be used for two different tasks: 1) To provide subtitles on a screen while the child is speaking, to support others in (learning to) understanding the child. 2) When integrated into the online-learning environment of the school, to help the child to fill out exercise sheets independently from the support teacher.

1.1 Dysarthric Child Speech

Diagnosis, aetiology, treatment and finding supportive solutions for speech disorders require multidimensional approaches including linguistic, psycholinguistic, medical and technical perspectives. We distinguish between speech disorders, which affect the physical production of sounds, and language disorders, which impact the ability to understand and use language structures. The presence of a speech disorder does thus not mean that there is no language system developed allowing to understand what is said and how it is said [5]. Various research on atypical speech production by children summarized in [5] show that the speakers intent to produce the right phonetic sounds, even when the listener can not identify them correctly. This emphasizes the cognitive nature of speech impairments in children, where the underlying system is intact and comprehensive, but the execution may vary significantly from typical speech.

Dysarthria constitutes a type of speech disorder where difficulties arise from impaired muscular control over the speech mechanisms due to neurological disorders caused in the central or peripheral nervous system (e.g., by conditions like Parkinson's disease, multiple sclerosis, stroke). Childhood dysarthria occurs as a result of neurological disorders that arise in early childhood, when motor speech abilities have not yet fully developed. Various components of speech production can be affected, such as phonation, articulation, respiration and prosody [5,6]. In phonation for example, the speaker might show a reduced vowel space, since the articulators never reach their intended positions, known as target undershoot. This may lead to a lower level of contrastivity between phones and a high variability in phonemes [5]. It is estimated that 50.000 children and adolescents in Germany are affected by dysarthria [6]. Depending on the neurological disorder causing the dysarthria, its severity, therapy applied and the child's developmental stage, the produced dysarthric speech and its characteristics are highly individual.

The symptoms of childhood dysarthria rarely occur in isolation. Co-morbidity is very common, with over half of the children with speech impairments also having language problems [5]. Although developmental speech impairments in children and acquired impairments in adults are intersectional, not much research in one field has been applied to the other so far [5,6]. This on the one hand leads to child speech not being included in standard research on dysarthria explicitly. There are symptoms of adult dysarthria that are not common in childhood dysarthria , e.g., pauses or iterations [6]. On the other hand, both speech development and neurological disorders have to be considered jointly. Finally, there are characteristics in child speech that stem from typical development, such as short respiration cycles due to physiological factors (e.g., small lungs or slow speech rate due to cognitive-linguistic development). These coincide with symptoms of dysarthria and their distinction may not always be possible [6].

The type of dysarthria of the child participating in the present study is medium-to-severe [1] ataxic dysarthria. It is characterized by both slow articulation and speech rates with long pauses in unusual locations in the sentence on the one hand, and seamless joining of words on the other. Another frequent characteristic is the insertion of vowels, with schwa (ə) being the most frequent. The child is on a reading level normative for its age and does not have a language impairment. The child speaks Austrian German.

[1] No official severity assessment has been made, but this is based on the speech therapist's judgement.

1.2 ASR for Pathological Speech

State-of-the-art ASR systems, such as OpenAI's Whisper [1] and Meta's Wav2Vec [2], have demonstrated remarkable performance across a wide range of speech tasks. Whisper is a transformer-based model, with its largest version,*large-v3*, trained on more than 5 million hours of labelled data. It has proven to be robust to noise and generalizes well across various datasets. In a zero-shot setting, meaning it has not been specifically fine-tuned for a particular task, it achieves WERs for German of 5.5% on Multilingual LibriSpeech, 6.4% on CommonVoice 9, and 11.2% on VoxPopuli [1]. Wav2Vec is based on self-supervised learning from raw audio representations that are unlabelled for pre-training and fine-tuned on a comparably small amount of labeled data. On the English LibriSpeech dataset, WERs of 17.3 % were achieved for only 1 hour of labelled training data and no language model. With 100 hours of labelled data and a transformer-based language model, WER of as low as 1.9 % are achieved [2].

Despite their strong performance under well-resourced conditions, both models experience significant performance degradation when labelled training data is scarce or when dealing with highly variable, atypical speech. For Austrian German conversational speech, zero-shot Whisper produces WERs ranging from 26% to over 63%, while Wav2Vec achieves WERs between 20% and 38% without a language model, and between 15% and 30% when using a 3-gram language model [7]. The challenges become even stronger for dysarthric speech, where WERs range from 70% to over 90%, depending on the severity of intelligibility impairments [8]. These results highlight the limitations of current ASR systems in low-resource and non-standard speech settings, underscoring the need for further adaptation techniques to improve robustness in such scenarios.

Developing effective ASR solutions for dysarthric child speech requires specialized models trained on representative datasets — yet such data is scarce, posing a major challenge. There exist some public datasets containing dysarthric speech. One set is the UA Speech dataset [9], containing speech by 19 American English speakers with cerebral palsy. The dysarthria diagnosis is mostly spastic or athetoid. Another set is the TORGO database [10], comprising 8 English speakers from Canada also affected by cerebral palsy or ALS, resulting in spastic, athetoid or ataxic dysarthria. The Dysarthric Expressed Emotional Database (DEED) [11] contains recordings of 4 British English dysarthric speakers affected by Parkinson's disease. There are currently no Austrian German or German databases with dysarthric speakers known to the authors. Also, all databases contain adult, mostly spastic or athetoid dysarthric speech, but no ataxic child speech. Thus, no public data can be effectively included in the training of our ASR system, and all data had yet to be collected. Summing up, for the development of an ASR system, especially the following characteristics of dysarthric child speech need to be considered:

1. For our child, the dysarthric speech follows a clear pattern that (after some accommodation) can be recognized by a human listener. This is the basic requirement needed for an ASR system to eventually be able to learn these patterns.
2. The high variation across dysarthric children requires data from the individual (if available in addition to other datasets).
3. With increasing age some of the characteristics will change over time or vanish completely (i.e., those stemming from typical development or those from speech therapy), which requires continuously learning ASR systems, so they develop with the child's development.

2. MATERIALS AND METHODS

Dysarthric speech can not be transcribed by lay persons, thus spontaneous sentences need to be transcribed by individuals who are familiar with the speaker's pronunciation. In our case, the parents and the speech therapist of the child created recordings and transcriptions over a longer period of time in advance of this work (i.e., *initial data set* (IDS)[2] The learnings from working with this set were put into the design of a recording tool, with which the subject can record new data at home. The tool and the *extended data set* (EDS) are described in subsec. 3.2. These data sets should contain many of the most frequently used lexemes, since our experiments showed that the ASR systems can more easily identify lexemes that were in the training data, as shown in sec. 3.1.

2.1 Initial Data Set

The *initial dataset* (IDS) includes single words, numbers, calculations, read stories, poems, jokes and sponta-

[2] **Ethical considerations:** IDS was provided to the authors of this paper after explicit written consent was obtained from the parents to use the recordings in anonymized form for research purposes. We respected the GDPR and the European Code of Conduct for Research Integrity.

Table 1. Types of utterances in the IDS and their amount, number of contained tokens and graphemes as well as total duration in minutes and word and grapheme rates.

utterance type	#utterances	#tokens	#graphemes	total duration	words/sec	graph./sec
spontaneous speech	186	784	3947	19.83 min.	0.66	3.32
read digits	122	431	2575	13.75 min.	0.52	3.12
read speech	403	1467	7675	35.36 min.	0.69	3.62
total	711	2682	14197	68.94 min.	0.65	3.43

neous conversations. We first sliced the audio into smaller chunks that can contain a word, a sentence or a coherent utterance and are not longer than 30 seconds each. Additionally, long pauses and other audible speakers were removed from the audio. In total, the IDS comprises 69 minutes of speech (i.e., 711 chunks with a total of 2682 word tokens, from which 1068 are unique lexemes). These units were sorted into three categories: 1) *Spontaneous speech* includes conversations over certain topics like the child's condition, favourite food or games and jokes. 2) *Read digits* contains single numbers and computational tasks. 3) *Read speech* contains stories and read words from the speech therapist's patho-linguistic diagnostic sessions.

For fine-tuning the models, IDS was split into a training split, containing 568 chunks, and a test split, containing 143 chunks. Between these splits, an overlap exists between the contained lexemes. In the split used for the initial fine-tuning and comparison of ASR-models 59% of the lexemes in the test set are also in the training set. This overlap is partly due to read stories and poems containing many reoccurring lexemes. Most of these are high frequency words such as the function words *der, die, das, und, mit, in*, the pronouns *sie, ich, es* and the verbs *ist, hat, war*. The most frequently occurring lexemes in the IDS show a similar pattern as those listed in the Austrian 1M sentences web dataset from the Wortschatz Leipzig corpus [12]. The fact that 40% of the lexemes are not present in the training set allows for an assessment of the system's generalization capability. However, it also indicates that, given the limited size of the dataset, a substantial number of lexemes commonly used in everyday language remain unknown to the system. To assess this problem in future work, a second dataset is created as described in sec. 3.2.

2.2 ASR Systems and Fine-tuning

For fine-tuning, all audio files were resampled to a uniform sampling rate of 16 kHz to ensure consistency across models. If data augmentation techniques were applied, the augmented audio data was incorporated into the dataset before training. Only training data was augmented. For *whisper* [13] and *wav2vec2-bert* -based [14] models, the audio was transformed into a mel-spectrogram representation using the respective feature extractors, whereas *wav2vec2* [14] was trained directly on raw audio waveforms. Additionally, all special characters (i.e., ! ? ") and similar symbols were removed from the text, and all text was converted to lowercase to standardize the training data. This preprocessing pipeline ensured compatibility between models and provided a uniform input representation for training and evaluation. Due to model size constraints, fine-tuning was only performed for *whisper-small*, *whisper-medium* and *wav2vec2(-bert)* models pre-trained with up to 300 million parameters. The larger models *whisper-large-v3* and 1 billion parameter *wav2vec2* models could not be fine-tuned.

Additionally to our own fine-tuning, models previously fine-tuned by the HuggingFace community on the *Common Voice German* datasets were used as a starting point for further fine-tuning. Common Voice will be abbreviated as cvXX-de, with *XX* standing for the version of the set, and *de* for the German subset. For *whisper-small/medium*, we used the models fine-tuned on cv11-de by HuggingFace-user bofenghuang [15]. For *wav2vec2-bert* the model fine-tuned on cv16-de by user sharrnah [16] were used. The *wav2vec2-xls-r-300m* model fine-tuned on cv11-de was provided by user aware-ai [17].

2.3 Data Augmentation

We implemented the following data augmentation techniques from SpeechBrain [18], while adjusting the parameters to the characteristics of dysarthria and the work environments of possible applications:

Clipping simulates audio clipping artifacts, which occur when the signal amplitude exceeds the dynamic range of the recording system. First, the waveform is normalized to a range between -1 and 1. Then, a predefined threshold is applied, setting all values beyond this threshold to the maximum or minimum value, effectively distorting the waveform. After clipping, the original amplitude is restored to maintain a comparable loudness level. This augmentation exposes the model to common distortions that may arise due to recording limitations or speaker volume fluctuations.

Time Drop involves randomly setting short temporal segments of the signal to zero, making parts of certain phonemes or syllables temporarily unavailable to the model. This simulates missing speech information due to brief interruptions, microphone dropouts, or speaker hesitations. By training on these incomplete signals, the model learns to infer missing speech content and improves its ability to handle irregular speech patterns.

Frequency Drop removes random frequency bands from the spectrogram to simulate real-world distortions (e.g., microphone artifacts or background noise masking certain frequencies). This forces the model to rely on broader spectral patterns rather than specific frequency components, thereby improving its robustness to environmental noise and speaker variability.

Speed Perturbation alters the playback speed of the audio by resampling the audio at sampling rates close to the original, effectively simulating variations in speech rate. Since the speech rate of our subject is slow and variable, speed perturbation was applied from 80% and 120% of the original speed, making the model more resilient to the variations observed in dysarthric speech.

Since augmentations effectively expand the fine-tuning dataset, comparisons with non-augmented models must control for the number of training steps. To ensure fairness, all models were fine-tuned until convergence, thereby accounting for this factor.

3. RESULTS

3.1 Performance Comparison of ASR Systems

The results of inference using different zero-shot and fine-tuned models are presented in tab. 2. Overall, the zero-shot models perform poorly, with WERs exceeding 100%, indicating a lack of generalization to dysarthric child speech. The high WERs of the *whisper* models are due to whisper containing a language model that can hallucinate in an infinite loop or it makes up several words fitting the duration of the audio.

Even with just 568 utterances from the IDS, fine-tuning significantly reduced the WER to approximately 54% for *whisper-small*. Applying augmentations as described in sec. 2.3 further improved performance to 46%. While previous fine-tuning on cv11 considerably improved WER compared to zero-shot models and slightly enhanced performance for models trained with augmented data, it did not yield improvements when used in combination with only unaugmented fine-tuning data. The best WER achieved was 46% for *whisper-small* and 33% for *whisper-medium*. For *wav2vec2-bert2.0* using the models previously fine-tuned on cv11 and cv16 shows a significant effect, lowering the WER for unaugmented data from 93% to 57% in comparison to *wav2vec2-bert2.0* without cv16. The best *wav2vec2-bert2.0* configuration yields a WER of 44%. In the case of *wav2vec2* it was not possible to generate meaningful output when only fine-tuning with our own data. Using augmentations again largely improved the WER to 50%, much more than for *whisper*.

To learn more about how to further improve ASR performance, we analysed WER separately for lexical items present in both the training and test splits and lexical items appearing only in the test split. For the *whisper-medium* model fine-tuned with four speed perturbation configurations (index 16 in Tab. 2), lexemes appearing in both sets had a WER of 16.5%, while unseen lexemes had a WER of 66.6%. This highlights two key aspects: first, the importance of acquiring a large and lexically diverse dataset for fine-tuning; second, the consistency of the child's speech patterns, as lexemes previously encountered during training were recognized more reliably by the ASR when presented again.

3.2 Recording Tool and Extended Data Set

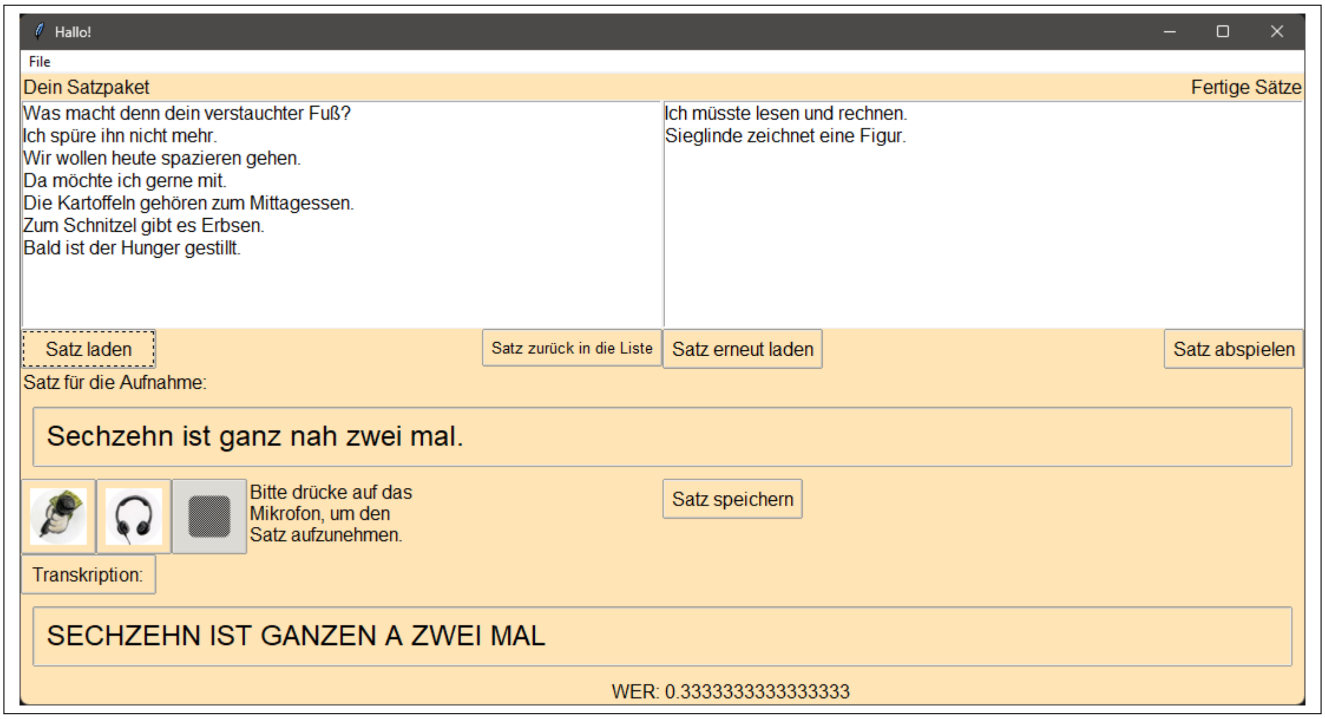
For the acquisition of additional data two factors are important: First, the recording process needs to integrate seamlessly into the daily routines of the family and can be done in a familiar environment (at home, at school). Second, the new dataset needs to be diverse and distinct from the IDS. To ensure both, we created a recording tool and provided sentence packages, continuously updated in the future recording process. Fig. 1 shows the tool's graphical user interface (GUI). It is designed in such a way that it is easy to use for both the family and the child itself. The child can choose which colour it wants the GUI to be in every time it starts the application. Big buttons with intuitive pictograms and a plain appearance make its use easy. Sentence packages are simple .txt-files, with one sentence or utterance per line and can thus also be created easily by the family or the speech therapist. They include 8-12 sentences each to keep recording session short. One sentence package is loaded into the list on the top left of the

Table 2. ASR performances on the *initial dataset's* (IDS) test-split. *ft-pre* refers to models already fine-tuned on healthy adult speech. *ft-IDS* refers to fine-tuning on IDS containing dysarthric child speech. Abbrevations: *cvXX*: Common Voice XX; *dc*: drop chunk; *df*: drop frequency; *cl*: clip; *spXX*: speed perturbation at XX %.

model	ft-pre	ft-IDS size	augmentations	epochs	WER	index
whisper-small	—	—	—	—	261 %	1
	—	568	—	10	54 %	2
	—	2272	dc; df; cl	10	52 %	3
	—	2840	sp80; sp90; sp110; sp120	10	47 %	4
	cv11	—	—	—	111 %	5
	cv11	568	—	10	54 %	6
	cv11	2272	dc; df; cl	10	52 %	7
	cv11	2840	sp80; sp90; sp110; sp120	10	47 %	8
	cv11	4544	dc; df; cl; sp80; sp90; sp110; sp120	10	46 %	9
whisper-medium	—	—	—	—	178 %	10
	—	568	—	10	45 %	11
	—	2840	sp80; sp90; sp110; sp120	10	40 %	12
	cv11	—	—	—	108 %	13
	cv11	568	—	10	42 %	14
	cv11	2272	dc; df; cl	10	36 %	15
	cv11	2840	sp80; sp90; sp110; sp120	10	35 %	16
	cv11	4544	dc; df; cl; sp80; sp90; sp110; sp120	10	33 %	17
whisper-large-v3	—	—	—	—	130 %	18
wav2vec2-xls-r-300m	cv11	—	—	—	104 %	19
	cv11	568	—	20	73 %	20
	cv11	2274	dc; df; cl	20	53 %	21
	cv11	4544	dc; df; cl; sp80; sp90; sp110; sp120	20	50 %	22
wav2vec2-bert-2.0	cv16	—	—	20	99 %	23
	—	568	—	20	93 %	24
	cv16	568	—	20	57 %	25
	cv16	2272	dc; df; cl	20	46 %	26
	cv16	4544	dc; df; cl; sp80; sp90; sp110; sp120	20	44 %.	27

GUI and sentences can then be chosen one by one. After recording, the child can listen to its own recorded speech using the play-back button and the recording can be overwritten or transcribed by the (currently best performing) ASR-model, which can easily be updated with future improved versions. The sentence and related audio are then saved to the list on the upper right of the GUI. Once the complete sentence package is finished, the text and audio files are saved to a folder, in a format designed to be ready for usage in fine-tuning. When testing the recording tool with the child, it began using the tool independently after a 5 min instruction phase, displaying immediate enthusiasm both in interacting with the tool and listening to its own voice. Incorrect ASR transcriptions were not frustrating; surprisingly they amused the child and showed to be an additional motivation to record additional material.

The tool currently includes includes 26 sentence packages, sourced from the reading material of the GRASS corpus [19]. Care was taken to ensure that all words were appropriate for children and that complex sentence structures were avoided. The reading material contains a total of 250 sentences, comprising 1367 word tokens. At the child's observed speaking rate, this will make up approximately 35 minutes of speech.



Dein Satzpaket

Was macht denn dein verstauchter Fuß?
Ich spüre ihn nicht mehr.
Wir wollen heute spazieren gehen.
Da möchte ich gerne mit.
Die Kartoffeln gehören zum Mittagessen.
Zum Schnitzel gibt es Erbsen.
Bald ist der Hunger gestillt.

Fertige Sätze

Ich müsste lesen und rechnen.
Sieglinde zeichnet eine Figur.

Satz für die Aufnahme:

Sechzehn ist ganz nah zwei mal.

Bitte drücke auf das
Mikrofon, um den
Satz aufzunehmen.

SECHZEHN IST GANZEN A ZWEI MAL

WER: 0.3333333333333333

Figure 1. GUI of the recording tool developed for collecting data easily from home or at school.

4. CONCLUSION

Our initial ASR results for dysarthric child speech are in line with the overall poor WERs reported in the literature for dysarthric speech [8]. Significant improvements were achieved through fine-tuning and data augmentation, demonstrating the effectiveness of targeted adaptation strategies, achieving best WERs of 33%. In comparison, for Austrian German adult speech, WERs lie between 13% and 47 % [7, 20]. Furthermore, we presented the first collection of Austrian German dysarthric child speech, with ongoing expansion enabled by the developed recording tool, allowing for continuous refinement of the training corpus and adaptation to changes due to typical development of the child.

We pursue several directions for future work on (1) resource generation and (2) improving ASR: (1) We plan to generate utterances automatically for new sentence packages using large language models (LLMs), based on the most frequent lexemes in Austrian German, utilizing resources such as the Wortschatz Leipzig corpus [12]. Another key future direction is speaker adaptation to produce larger datasets, independent from the recording capabilities of the child, leveraging techniques such as those outlined in [21] and [22]. (2) The speech therapist provided us the notes on the consistencies of phonetic processes. We plan to use these rules to create a knowledge-based lexicon that could be integrated in the ASR system following the hybrid approach recently presented by Perikh, A. et al. [23]. The lexicon will include statistical information on the frequency of phonetic processes (e.g., schwa insertion) in different phonetic contexts currently being investigated by Galovic, M. [24]. Further improvements will focus on adding a language model (LM) to *Wav2Vec2* and *Wav2vec2-Bert*, exploring both transformer-based and n-gram approaches. Training efforts may also be extended to larger models, including *whisper-large-v3* and *wav2vec2*-models pre-trained with 1 billion parameters, to assess performance gains with increased model capacity. To conclude, this paper shares valuable experiences with data collection for pathological child speech. Currently available ASR systems, as powerful as they may be on commercial tasks, they do not only perform poorly, but ”not at all” for pathological child speech. Since ASR systems are the ”bottle-neck” of assistive technology and online learning environments, our work highlights the necessity for putting more effort in collecting data and developing assistive technology directly together with those who need it.

5. REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[3] T. Schölderle, E. Haas, and W. Ziegler, *Dysarthrien bei Kindern*. Schulz Kirchner Verlag, 2020.

[4] D. McNaughton, T. Rackensperger, E. Benedek-Wood, C. Krezman, M. B. Williams, and J. Light, "A child needs to be given a chance to succeed: Parents of individuals who use AAC describe the benefits and challenges of learning AAC technologies," *Augmentative and Alternative Communication*, vol. 24, no. 1, pp. 43–55, 2008.

[5] J. S. Damico, N. Müller, and M. J. Ball, *The handbook of language and speech disorders*. Wiley Online Library, 2010.

[6] E. Haas, *Developmental courses of childhood dysarthria*. PhD thesis, LMU, 2021.

[7] J. Linke, B. C. Geiger, G. Kubin, and B. Schuppler, "What's so complex about conversational speech? a comparison of HMM-based and transformer-based asr architectures," *Computer Speech & Language*, vol. 90, p. 101738, 2025.

[8] L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," 2022.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Interspeech 2008*, pp. 1741–1744, 2008.

[10] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, Mar. 2011.

[11] L. Alhinti, S. Cunningham, and H. Christensen, "The dysarthric expressed emotional database (DEED): An audio-visual database in British English," *PLOS ONE*, vol. 18, no. 8, p. e0287971, 2023.

[12] D. Goldhahn, T. Eckart, and U. Quasthoff, "Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages," in *Proc. of LREC*, 2012.

[13] OpenAI, "Whisper Model (small, medium, large-v3)." `https://huggingface.co/openai`, 2023. Release: openai-whisper==20231106.

[14] Facebook, "Wav2vec2 Model (base-960, xls-r-300m, bert2.0)." `https://huggingface.co/facebook`, 2022. Release: facebook-wav2vec2==20221114.

[15] bofenghuang, "Whisper models fine-tuned on common voice 11." `https://huggingface.co/bofenghuang`, 2023. Release: whisper-cv11-de==20221227.

[16] sharrnah, "Wav2vec2-bert model fine-tuned on common voice 16." `https://huggingface.co/sharrnah/wav2vec2-bert-CV16-de`, 2024. Release: wav2vec2-bert-cv16-de==20240211.

[17] aware-ai, "Wav2vec2 model fine-tuned on common voice 11." `https://huggingface.co/aware-ai/wav2vec2-xls-r-300m-german-cv11`, 2022. Release: wav2vec2-cv11-de==20220920.

[18] M. Ravanelli, T. Parcollet, and P. P. et al., "Speech-Brain: A general-purpose speech toolkit," 2021. arXiv:2106.04624.

[19] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of Read And Spontaneous Speech," in *Proc. of LREC*, pp. 1465–1470, 2014.

[20] J. Linke, P. N. Garner, G. Kubin, and B. Schuppler, "Conversational speech recognition needs data? Experiments with Austrian German," in *Proc. of LREC*, pp. 4684–4691, 2022.

[21] M. K. Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, and J. Černocký, "Speaker adaptation for wav2vec2 based dysarthric ASR," 2022.

[22] H. Wang, Z. Jin, M. Geng, S. Hu, G. Li, T. Wang, H. Xu, and X. Liu, "Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *Proc. of ICASSP*, pp. 12311–12315, 2024.

[23] A. K. Parikh, L. ten Bosch, and H. van den Heuvel, "Ensembles of hybrid and end-to-end speech recognition.," in *Proc. of LREC-COLING*, pp. 6199–6205, 2024.

[24] M. Galovic, "Investigating self-supervised pretraining frameworks for pathological speech recognition," in *Proc. of 3rd Graz-Vienna Speechworkshop. Connecting with health sciences*, pp. 17 – 18, 2025.