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

Automatic detection of speech sound disorders in German-speaking children: augmenting the data with typically developed speech

Darline Monika Marx[1], *Marco Matassoni*[2], *Alessio Brutti*[2]

[1]Free university of Bozen-Bolzano, Italy
[2]Augmented Intelligence Center, Fondazione Bruno Kessler, Italy
`DarlineMonika.Marx@student.unibz.it, matasso@fbk.eu, brutti@fbk.eu`

Abstract

Speech Sound Disorders (SSD) are common among children, affecting their academic, social, and emotional development. Traditional diagnostic methods are based on speech-language pathologists, making them resource-intensive. Due to the global shortage of experts and increasing demand, exploring deep-learning tools is crucial. Adapting a multi-task framework to fine-tune a pre-trained multilingual Wav2Vec model, this study tackles Automatic Speech Recognition and SSD classification for German children using a custom dataset. We show that incorporating public out-of-domain datasets improves robustness and generalizability. Interestingly, combining pathological and typical speech data with mispronunciations benefits the performance in terms of speech recognition and SSD detection. Finally, we investigate a two-step training of the model that further improves the overall performance.

Index Terms: speech sound disorders, children speech, classification, multi-task learning, speech recognition

1. Introduction

Speech Sound Disorders (SSDs) are a common reason for speech therapy, affecting approximately 16% of kindergarten and preschool-aged children in Germany [1]. SSDs [2] involve difficulties in acquiring or using speech sounds appropriate for a child's age and are classified into organic and functional types.

Early diagnosis and intervention are crucial to preventing long-term effects on literacy, education, and social interactions. Without treatment, SSDs can hinder reading and spelling skills, limiting academic and career opportunities. Socially, SSDs can cause peer difficulties, stigmatization, and emotional challenges such as low self-esteem and social withdrawal [3, 4].

Speech-language pathologists use standardized assessments to diagnose SSDs, but the process is time-consuming and impacted by a shortage of professionals. Advances in artificial intelligence and automated speech processing offer potential solutions to enhance diagnosis and therapy efficiency, improving accessibility and scalability in clinical and educational settings. The shortage of such experts has driven interest in automated tools to diagnose and treat SSDs: these tools improve accessibility and affordability, streamline diagnostic processes, and support early intervention. Recent developments include AI-driven speech analysis software and interactive therapy platforms [5]. Tools like Vocametrix [6] provide objective speech assessments, while Tabby Talks [7] and Buddy [8] assist children with specific speech disorders through phonological analysis and interactive exercises. However, many systems lack robust Automatic Speech Recognition (ASR) or struggle with the variability of disordered child speech. In Germany, applications like neolexon Artikulations-App [9] face similar limitations. A key challenge remains the scarcity of large, annotated datasets, which restricts model generalizability. To address this, pre-trained deep acoustic models using self-supervised learning offer a promising approach, improving speech disorder detection and supporting integration into clinical and educational settings.

Deep acoustic models have revolutionized ASR by boosting transcription accuracy, even in resource-constrained languages [10]. Beyond ASR, they support speech classification tasks, including speaker identification, emotion detection, and accent recognition, enabling applications such as voice assistants and emotion-aware systems [11, 12]. These models are also used to detect pathological speech patterns, such as voice disorders, aphasia, and stuttering. Studies show that pre-trained models like wav2vec2 [10, 13] and HuBERT [14], combined with domain-specific data, improve classification performance [15]. Cross-lingual approaches further enhance aphasia detection [16], while fine-tuned models improve stuttering classification accuracy [17]. Deep acoustic models are also applied in SSD detection, highlighting their growing role in speech disorder analysis [5].

This work investigates a multi-task learning approach for detecting SSDs in German-speaking children. We build upon [18], a framework that combines ASR and pronunciation classification, using the pre-trained wav2vec2 model to handle pathological speech. In particular, we explore how publicly available SSD data as well as typically developed speech can be employed to enhance the model performance and robustness. To do so, a collateral contribution is the creation of a custom dataset of speech recordings from German-speaking children with SSD, addressing data scarcity. Finally, we investigate architectural improvements, such as a two-step training process and selective classification-head layers, to optimize model performance.

2. Methodology

This work investigates the use of Wav2Vec2.0 [10], a self-supervised speech encoder designed to provide robust and effective speech representations, in a multi-task training framework to handle both ASR and pronunciation classification tasks within a unified architecture, following the paradigm in [18]. The architecture includes a convolutional feature encoder and 24 Transformer layers, to extract the speech representations, that can then be fine-tuned for specific tasks. The model features two distinct output heads: one for ASR, which generates word-level transcriptions, and another for pronunciation classification, which evaluates speech quality. The training process integrates two loss functions, CTC loss for ASR and cross-entropy loss for pronunciation classification, preventing catastrophic forgetting.

10.21437/Interspeech.2025-1996

The ASR head is trained with Connectionist Temporal Classification (CTC) loss, which allows alignment-free training by mapping input sequences X to transcriptions Y of variable length. The CTC loss is computed as:

$$\mathcal{L}_{CTC} = -\sum_{(X,Y)} \log \sum_{\pi \in A(Y)} P(\pi|X)$$

where X is the input speech representation, Y is the target transcription, $A(Y)$ represents all possible alignments of Y, π is a sequence in the alignment set. For the pronunciation classification, the cross-entropy loss is applied to measure the difference between predicted class probabilities and ground truth labels:

$$\mathcal{L}_{CE} = -\sum_{i} y_i \log \hat{y}_i$$

where y_i is the true class label and $\hat{y}_i$ is the predicted probability for class i. Following [18], both loss functions are equally weighted and summed:

$$\mathcal{L} = \mathcal{L}_{CTC} + \mathcal{L}_{CE}$$

Figure 1 shows the architecture with the two classification heads and the related losses. Experiments in [18] have already demonstrated the efficacy of the multi-task approach; therefore, we do not ablate the loss combination in this study.

2.1. SSD classification head

The Wav2Vec2.0 encoder features a series of transformer layers hierarchically structured. The ASR or speech classification head is typically placed at the top-end of the stack. However, it is commonly established that lower layers approximately focus on acoustic features, intermediate layers on phoneme and word representations, and higher layers on semantic and contextual information. Therefore, as explored in [18], it makes sense to assume that a pronunciation classification task must rely more on acoustic features. Hence, in this research, we also investigate optimizing the placement of the SSD classification head within these layers to improve performance.

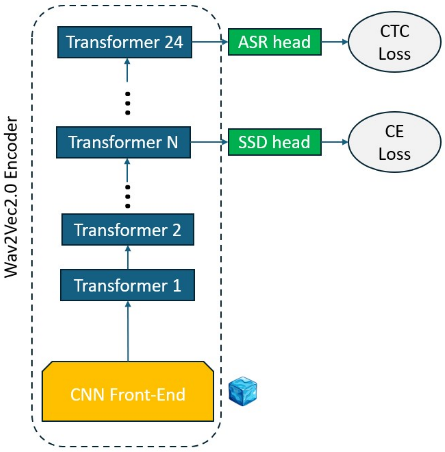

Figure 1: *Architecture of the multi-task Wav2Vec2 framework. Adapted from [18]. The ASR head is placed at the end of the stack and contributes to the CTC loss. The SSD classification head is placed after an intermediate transformer layer and contributes to the CE loss.*

2.2. Encoder pre-fine-tuning

Pre-trained speech encoders are generally trained on typical adult speech. Therefore, the multi-task training is in charge of both adapting the acoustics of the model to children's speech [19] and of optimizing the speech representation for the SSD tasks. The former, however, does not actually necessitate pathological speech but can leverage typically developed children's speech. Therefore, we propose a two-step training strategy. In the first step, the pre-trained ASR model is fine-tuned on correctly pronounced child speech data to improve its transcription of child speech, in a sort of pre-fine-tuning. In the second step, the pre-fine-tuned encoder is employed in the standard multi-task framework, training it on both ASR and pronunciation classification tasks. This two-step approach aimed to improve overall model performance by specializing the ASR component for children's speech before performing both tasks.

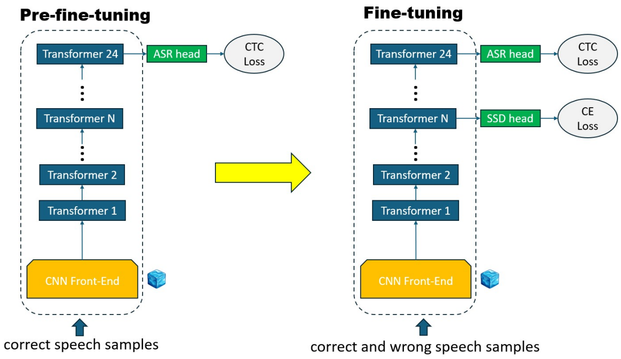

Figure 2: *Schematic overview of the encoder pre-fine-tuning strategy. The german Wav2Vec2.0 encoder is fine-tuned on correctly pronounced children's speech. Then, the standard multi-task method is applied using both mispronounced and correctly pronounced audio.*

3. Experimental Setup

This study employs three corpora of German child speech for training and evaluation: Neumann/Fox-Boyer (NFB), Fox-Boyer (FB), and a newly collected dataset (DM). These datasets contain speech recordings from children with and without SSD. To ensure consistency in model training and evaluation, a speaker-dependent split was applied to all datasets, which means that all speech samples from a single child remain within the same partition. Table 1 provides an overview of the dataset statistics and class distributions.

Table 1: *Dataset statistics and phonological error distribution across NFB, FB, and DM corpora.*

Dataset	# spk.	# words	Mispronounced	Correct
NFB	29	2,678	47.0%	53.0%
FB	32	2,977	16.1%	83.9%
DM	35	2,174	39.6%	60.4%

3.1. Neumann/Fox-Boyer and Fox-Boyer datasets

We employed two publicly available PhonBank corpora [20]. The **Neumann/Fox-Boyer** (NFB) corpus consists of recordings from 29 German-speaking children (aged 3 to 10 years) diagnosed with SSD. The **Fox-Boyer** (FB) corpus contains speech from 32 typically developing children (aged 2 to 9 years). Although children in the FB corpus are typically developing speakers, their speech may still contain mispronunciations due to ongoing phonological acquisition. As children gradually establish an underlying representation of speech sounds, they may omit, substitute, or modify certain phonemes. These patterns, known as phonological processes, reflect age-appropriate development [21].

Both corpora consist of speech samples elicited through the PLAKSS picture-naming task, which includes 96 target words that cover all vowels, consonants, and consonant clusters in German, with variations in word length and stress patterns [1]. Each corpus includes timestamps and phonetic transcriptions for both the target word and the child's actual utterance [22]. For this research, we used a binary classification scheme, distinguishing between correctly pronounced and mispronounced words. While a three-level classification system was initially developed (differentiating between words affected by a single phonological process and those exhibiting multiple phonological processes), our analysis in this paper is exclusively based on the binary classification task.

3.2. Custom dataset

Additionally, a custom dataset (DM) was collected from 35 children (aged 3 to 7 years) diagnosed with SSD, recorded between August and November 2024 in speech therapy practices in Stuttgart (Baden-Württemberg) and Hanau (Hesse). The dataset includes 19 male and 16 female children. The PLAKSS picture-naming test was used for all assessments. Depending on the diagnostic context, 16 recordings employed the screening version (30 items), while 19 recordings used the full version (96 items). The raw therapy recordings were manually segmented in Praat [23], isolating individual word-level utterances. Each segment was transcribed and labelled, following the same classification schemes applied to the Phonbank corpora (see Figure 3 for an example of manual transcription and labelling). Since this dataset contains personally identifiable audio recordings of children, it will not be publicly released due to ethical and data protection concerns.

3.3. Dataset Combinations

We consider two approaches to augment our custom dataset DM with publicly available data. In both cases, our evaluation is carried out on the test sets of DM and NFB.

- *COMBI* considers only data with pathological speech as it merges the training sets of NFB and DM.
- *COMBI+* extends COMBI by incorporating selected samples from FB, specifically recordings from nine speakers where at least 25% of the words were mispronounced. This ensures that additional data from typically developing children do not introduce excessive class imbalance while still contributing to phonological variability.

3.4. Implementation Details

For this study, we employ a Wav2vec2 model optimized for German speech[1]. Training was performed with a batch size of 16, a learning rate of 5e-5, and 50 epochs. A warm-up ratio of 0.1 was applied, and Unweighted Average Recall (UAR) was used as the evaluation metric for the best model. A sampling rate of 16 kHz was used for all datasets. The classification head was initially placed at layer 24, but additional experiments evaluated placements at layers 17 and 20 to optimize classification performance. We employed the code of [18][2].

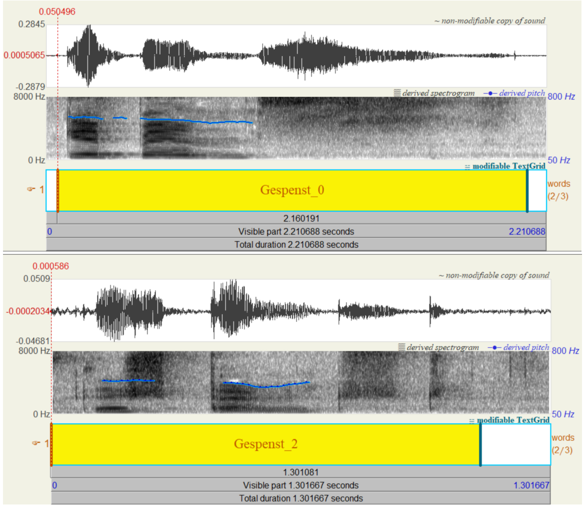

Figure 3: *Example of manual transcription and labelling in the DM dataset. The spectrograms show the word Gespenst (ghost) produced incorrectly (top) and correctly (bottom), with corresponding TextGrid annotations.*

4. Results

4.1. Evaluation Metrics

The model's performance was evaluated using Word Error Rate (WER) and Character Error Rate (CER) for ASR, and Accuracy (ACC), Unweighted Average Recall (UAR), and F1-Score for classification. WER measures transcription errors at the word level, while CER evaluates character-level mistakes. Both were calculated only for correctly pronounced words, excluding mispronounced samples. Accuracy provides a general classification overview but can be biased in imbalanced datasets. UAR ensures a balanced evaluation across classes by averaging recall values, making it particularly relevant for SSD classification. F1-Score, the harmonic mean of precision and recall, accounts for both false positives and false negatives, offering a comprehensive assessment.

4.2. Baseline Performance

The baseline performance of the multi-task model was evaluated by training separately on the NFB and DM datasets. The model achieved 76.5% UAR, 40.1% WER, and 18.6% CER on NFB, while on DM, it obtained 75.2% UAR, 34.0% WER, and 16.3% CER. These results are summarized in the top rows of Table 2.

4.3. Effect of Dataset Combination

To assess the impact of increased training data, the model was trained on the COMBI dataset, which merges the training sets

---

[1] `https://huggingface.co/facebook/wav2vec2-large-xlsr-53-german`

[2] `https://github.com/aalto-speech/multitask-wav2vec2`

of NFB and DM. This was further extended to COMBI+, incorporating selected FB samples. Training on COMBI led to improved performance, achieving 77.8% UAR on DM and 77.4% on NFB, while reducing WER to 26.5% for DM and 28.3% for NFB. Even if the acoustic quality of the two datasets is different, their combination improves performance on both datasets substantially, in particular for ASR metrics. Very interestingly, **the inclusion of additional typically developed speech data in COMBI+ further enhanced performance**, particularly for NFB test set, reaching 80.4% UAR and 20.9% WER. Using more data probably helps the model adapt to the children's speech, and the mispronunciations, although not pathological, seem to help detect correct speech segments. A detailed comparison of these results can be found in Table 2.

Table 2: *Performance metrics across individual datasets (NFB, DM) and dataset combinations (COMBI, COMBI+).*

Train	Test	WER (%)	CER (%)	ACC (%)	UAR (%)	F1 (%)
NFB	NFB	40.1	18.6	76.5	76.5	76.5
DM	DM	34.0	16.3	77.2	75.2	75.1
COMBI	NFB	28.3	14.5	77.3	77.4	77.1
	DM	26.5	12.0	78.3	77.8	77.6
COMBI+	NFB	**20.9**	**11.0**	**80.4**	**80.4**	**80.4**
	DM	**22.6**	**11.5**	**80.8**	**78.7**	**78.9**

4.4. Classification Head Placement

Following the results reported in [18], the effect of classification head placement within the Transformer layers was analyzed by testing configurations from layer 16 to layer 24. The impact of classification head selection is illustrated in Figure 4. The best performance for DM was observed at layer 17, with 79.7% UAR, while NFB benefited from placement at layer 20, achieving 79.4% UAR. WER was also lowest when placing the classification head at layer 17 for DM (22.4%) and layer 20 for NFB (22.0%).

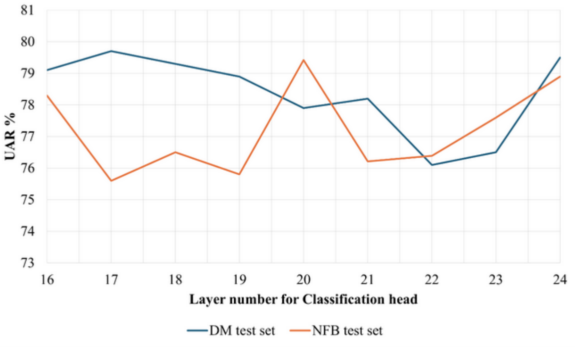

Figure 4: *Unweighted Average Recall (UAR) across Transformer layers 16 to 24 for the classification head, comparing performance on the DM and NFB test sets.*

4.5. Encoder pre-fine-tuning.

Finally, Table 3 reports the performance of the encoder pre-fine-tuning. The two-step training approach was designed to enhance pronunciation classification by first fine-tuning the ASR model exclusively on correctly pronounced word segments from NFB, FB, and DM before integrating it into the multi-task framework. Based on the findings of the classification head placement analysis, the experiment incorporated the optimal classification head layers identified in the previous analysis. By placing the classification head at layer 20 for NFB and layer 17 for DM, the model achieved improvements over the original single-step framework. For NFB, two-step training increased UAR from 80.4% to 81.7% and reduced WER from 20.9% to 19.8%. Similarly, for DM, UAR improved from 78.7% to 80.7%, while WER decreased from 22.6% to 20.0%.

Table 3: *Performance metrics for two-step training; optimal classification head uses layer 20 for NFB and layer 17 for DM (see Section 4.4). These models are fine-tuned only on COMBI+.*

Train	Test	WER (%)	CER (%)	ACC (%)	UAR (%)	F1 (%)
COMBI+	NFB	21.1	11.9	79.3	79.4	79.3
	DM	20.2	10.6	81.0	79.5	79.4
+ optimal class head layer						
COMBI+	NFB	**19.8**	**10.5**	**81.7**	**81.7**	**81.6**
	DM	**20.0**	11.3	**82.7**	**80.7**	**81.0**

5. Conclusion

The work investigates a multi-task learning framework combining ASR and pronunciation classification to detect SSDs in German-speaking children. By leveraging the pre-trained wav2vec2 model, it demonstrated the feasibility of deep learning for pathological child speech and investigated architectural modifications to optimize performance.

Experimental findings highlight the impact of two-step training and classification head layer selection on model performance. Fine-tuning on correctly pronounced speech in a two-step approach improved both ASR and classification. Additionally, placing the classification head at intermediate Transformer layers yielded better performance than the final layer, reinforcing the need for dataset-specific adaptations.

A central contribution of this study was the DM dataset, a custom collection of speech recordings from children with SSDs. Despite its limited size, incorporating publicly available corpora (NFB, FB) enhanced model robustness, demonstrating that combining pathological and typically developing speech data mitigates data scarcity challenges.

While promising, the approach is constrained by dataset size, reliance on manual segmentation, and simplified phonological labelling. Future work should explore larger datasets, alternative model architectures, and a more detailed classification approach.

This research provides a solid foundation for the development of automated SSD detection tools in German-speaking children, advancing automated speech disorder diagnosis with implications for clinical practice and future research.

6. References

[1] A. Fox-Boyer, “Aussprachestörungen im deutschen,” in *Handbuch Spracherwerb und Sprachentwicklungsstörungen Kindergartenphase*, A. Fox-Boyer, Ed. München: Elsevier, 2014.

[2] American Speech-Language-Hearing Association, “Speech sound disorders: Articulation and phonology,” 2024.

[3] J. McCormack, S. McLeod, L. J. Harrison, and L. McAllister, “The impact of speech impairment in early childhood: Investigating parents’ and speech-language pathologists’ perspectives using the icf-cy,” *Journal of Communication Disorders*, vol. 43, no. 5, pp. 378–396, 2010.

[4] J. L. Preston, M. Hull, and M. L. Edwards, “Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders,” *American Journal of Speech-Language Pathology*, vol. 22, no. 2, pp. 173–184, 2013.

[5] Z. Brahmi, M. Mahyoob, M. Al-Sarem, J. Algaraady, K. Bousselmi, and A. Alblwi, “Exploring the role of machine learning in diagnosing and treating speech disorders: A systematic literature review,” *Psychol Res Behav Manag*, pp. 2205–2232, 2024.

[6] Vocametrix, “Vocametrix platform,” 2024, retrieved December 29, 2024, from https://platform.vocametrix.com/.

[7] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, “Tabby talks: An automated tool for the assessment of childhood apraxia of speech,” *Speech Communication*, vol. 70, pp. 49–64, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639315000382

[8] P. Ramamurthy and T. Li, “Buddy: A speech therapy robot companion for children with cleft lip and palate (cl/p) disorder,” in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’18. Association for Computing Machinery, 2018, p. 359–360.

[9] H. Jakob and M. Späth, “Sprachtherapeutische apps am beispiel neolexon: Herausforderungen beim zugang in die versorgung und chancen für therapeuten und patienten,” *Sprache· Stimme· Gehör*, vol. 45, no. 01, pp. 17–21, 2021.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, 2020, pp. 12 449–12 460.

[11] B. W. Schuller, “Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends,” *Commun. ACM*, vol. 61, no. 5, p. 90–99, Apr. 2018. [Online]. Available: https://doi.org/10.1145/3129340

[12] J. Zhang, Y. Peng, V. T. Pham, H. Xu, H. Huang, and E. S. Chng, “E2E-based Multi-task Learning Approach to Joint Speech and Accent Recognition,” in *Proceedings of Interspeech*. International Speech Communication Association, 2021, pp. 1519–1523.

[13] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, “wav2vec2-based speech rating system for children with speech sound disorder,” in *Interspeech*, 2022, pp. 3618–3622.

[14] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Rahman Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[15] D. Ribas, M. A. Pastor, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, “Automatic voice disorder detection using self-supervised representations,” *IEEE Access*, vol. 11, pp. 14 915–14 927, 2023.

[16] G. Chatzoudis, M. Plitsis, S. Stamouli, A. Dimou, N. Katsamanis, and V. Katsouros, “Zero-shot cross-lingual aphasia detection using automatic speech recognition,” in *Proc. of Interspeech*, 2022, pp. 2178–2182.

[17] S. Bayerl, D. Wagner, E. Noeth, and K. Riedhammer, “Detecting dysfluencies in stuttering therapy using wav2vec 2.0,” in *Proc. of Interspeech*, 2022, pp. 2868–2872.

[18] Y. Getman, R. Al-Ghezi, T. Grosz, and M. Kurimo, “Multi-task wav2vec2 serving as a pronunciation training system for children,” in *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 36–40.

[19] P. G. Shivakumar and S. S. Narayanan, “End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study,” *Comput. Speech Lang.*, vol. 72, 2021.

[20] Y. Rose and B. MacWhinney, “380the phonbank project: Data and software-assisted methods for the study of phonology and phonological development,” in *The Oxford Handbook of Corpus Phonology*. Oxford University Press, 05 2014.

[21] B. Dodd, *Differential diagnosis and treatment of children with speech disorder*. London: Whurr, 1995.

[22] A. Fox-Boyer, *PLAKSS-II Psycholinguistische Analyse kindlicher Aussprachestörungen (psycholinguistic analysis of childhood speech sound disorders)*, 2nd ed. Frankfurt: Pearson Assessment, 2014.

[23] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer,” 2024.