

AUTOMATIC DETECTION OF DISFLUENCIES IN L1 AND L2 CHILD SPEECH

Martha Schubert¹, Valentin Kany^{1,2}

Saarland University,

*¹Language Science and Technology; ²German as a Foreign and Second Language
masc00114@stud.uni-saarland.de, valentin.kany@uni-saarland.de*

Abstract: Language Proficiency Assessments (LPAs) for preschool children often rely on manual evaluation methods that suffer from subjectivity and low inter-rater reliability. Moreover, current approaches typically overlook speech fluency which is an important indicator of language proficiency and can be linked to overall linguistic development. To address these limitations, we propose a more consistent and comprehensive LPA framework that incorporates automatic speech fluency assessment through disfluency detection. Disfluencies such as repetitions, repairs, restarts, partial words and filler particles significantly influence perceived fluency and thus serve as key indicators for modeling. Using spontaneous speech data collected from 167 German-speaking preschoolers via the “Wuschel-App,” we annotated a subset of utterances with fine-grained disfluency labels and fine-tuned a BERT-based token-level sequence labeling model, inspired by Romana et al.’s [1] architecture. Despite pronounced class imbalance in the dataset, our best model achieves a token-level accuracy of 96.0% and a macro F1 score of 65.7%, substantially outperforming the majority-class baseline (accuracy = 56.86%). Error analysis reveals that misclassifications primarily affect the rarest disfluency categories. We anticipate that expanding the annotated corpus will improve performance on these underrepresented classes, enabling finer-grained modeling and, ultimately, a robust text-signal combined fluency-aware LPA system.

1 Background

Language proficiency plays an important role in children’s social integration and educational success [2]. In German child daycare centres (kindergartens), heterogeneity in children’s language backgrounds has increased over recent years, resulting in highly fluctuating language proficiency levels among preschool children [3]. Thus, Language Proficiency Assessments (LPAs) are conducted during preschool age to ensure language learning support for children lagging behind in language development at an early stage. Early intervention helps these children catch up. This way, all children should reach a comparable level of language proficiency by the time they enter school, providing more equal opportunities in education.

Typical LPAs for preschool children in Germany focus on vocabulary, grammar [4, 5], and morphology [6]. Alongside these factors, speech fluency has proven to be a significant indicator of language proficiency in several studies (e.g. [7, 8, 9]), yet it is insufficiently considered by the LPAs practiced in Germany. Therefore, we plan on extending WUSCHEL [10], an innovative game-based LPA-method with automatic assessment, by adding automatic speech fluency scoring. In previous studies, we collected data utilising WUSCHEL and manually annotated parts of it on various types of disfluencies. These phenomena break the perception of fluent speech and are thus useful to assess speech fluency. The annotations were used to extract fluency-related

features that build the basis of individual fluency profiles, which we developed as a way to analyse a child’s speech fluency [11]. However, to align with WUSCHEL’s concept as an automatic LPA, the annotation process needs to be fully automated to immediately generate these fluency profiles after recording a child. In this study, we aim to take a first step towards the automatic annotation of disfluencies in spontaneous child speech by detecting disfluency types from an orthographic transcript.

2 Previous Work

2.1 Disfluency Detection in General

Automatic disfluency detection in children’s speech remains challenging and underexplored. Venkatasubramaniam et al. [12] proposed an end-to-end attention-based model for word-level disfluency detection and classification in children’s read speech, first distinguishing fluent from disfluent utterances ($F1 = 42.8\%$) and then categorising disfluencies into types such as repetitions, mispronunciations, false starts, skips, prompts, and tracking errors ($F1 = 27.6\%$), highlighting the difficulty of fine-grained typing.

Tran et al. [13] showed that a model trained on adult speech generalises reasonably well to child speech ($F1 = 77\%$), but their scheme only distinguishes fluent, reparandum, repair, or both—lacking granularity. Similarly, Proença et al. [14] used a Hidden Markov Model with Mel-Frequency Cepstral Coefficients and Voice Activity Detection to detect repetitions and pre-corrections, reporting a 15% miss rate and 5% false alarm rate, but omitted other disfluency types like filler particles or revisions.

Liu et al. [15] leveraged wav2vec 2.0 and Convolutional Neural Network (CNN)/ Transformer models to detect disfluencies—including repetitions, prolongations, interjections, and blocks—across multiple languages, yet their system was trained solely on adult speech, limiting applicability to children. Yildirim and Narayanan [16] combined audio and visual cues to detect disfluency boundaries in spontaneous child speech, achieving 82.1% accuracy with multimodal fusion, but only performed binary fluent/disfluent classification.

Other work focuses solely on filler particles such as "uh" and "um": An et al. [17] used frame-level Support Vector Machines (89.3% accuracy on ComParE), while Chatziagapi et al. [18] fused CNNs and Recurrent Neural Networks on Automatic Speech Recognition output (73.9% accuracy), with better results on manual transcripts. Additional approaches rely on acoustic-prosodic cues (e.g., stable formants, flat pitch, energy dynamics) but are often rule-based or language-specific.

Collectively, existing methods are limited to read speech, adults, binary labels, or few disfluency types—and rarely address the full spectrum of disfluencies in spontaneous child speech, especially across L1 and L2 learners. Our work bridges this gap by leveraging linguistic context via a BERT-based sequence labeling model trained on richly annotated child speech data, inspired by Romana et al. [1].

2.2 Reimplemented Model

Romana et al. [1] trained their model on the Switchboard corpus [19], a dataset comprising 260 hours of English adult telephone conversations. While this resource provides a valuable foundation for disfluency detection in spontaneous speech, it differs substantially from our target domain (German child speech) in language, age group, and recording context. Consequently, we re-implement the overall modeling framework from scratch, adapting it to our data and task. Our approach is nonetheless inspired by Romana et al.’s architecture, as we also employ a

linguistic model based on BERT [20]. However, while [1]’s recordings are transcribed automatically using Whisper, we test with manually transcribed data. [1] also include an acoustic model built upon WavLM [21]. A text-signal combined variant that fuses both linguistic and acoustic signals is also planned. Due to ongoing development of the acoustic component, results for this integrated model will be reported in future work.

3 The Dataset

3.1 Acquisition method

Using the "Wuschel-App" [10], a game-based elicitation method, we collected spontaneous speech data from 167 children (aged 4-6 years) in 30 kindergartens in Saarland, Germany. The game presents a story consisting of 28 coherent scenes, in which the main character "Wuschel", a dog, requires the child’s help to progress. In each scene, Wuschel asks two questions, the second one being a follow-up question to give the child another chance to answer. The child’s answers to these questions are recorded using the built-in microphone of the iPad (9th generation) on which the game is played. While the use of an external microphone would result in higher audio quality, it could create an uncomfortable situation for the child and affect the naturalness of our speech data. The game follows a Wizard-of-Oz principle: While the child playing the game believes Wuschel acts autonomously, the game is directed by a confederate of the experimenters via a second app. This allows the confederate to decide when Wuschel poses his questions and adjust the timing according to the child’s responses. This further enhances the naturalness of the dialogue and the recorded speech data. Each kindergarten provided a separate room for us to collect our data. During the recording sessions, only the child, the confederate, and one member of the kindergarten’s staff were present in the room. An average recording session lasted approximately 30 minutes which we consider reasonable for the children, as reflected by a relatively low dropout rate of 2.5%.

3.2 Data

The game’s setup results in a total of 56 (28 scenes with 2 questions each) recorded segments of speech per child. The average duration of these segments is 8.46 seconds (including pauses) and 3.23 seconds of articulation time (excluding pauses). These rather short segments of coherent speech result from the question-answer-like structure of the game. Usually, the children briefly explain to Wuschel what to do or describe what happened in the scene and wait for Wuschel’s reaction. However, the second question of each scene tends to elicit a more elaborate answer as Wuschel asks for further clarification. Our data includes both speech from native (L1) and non-native (L2) speakers of German. Out of the 167 recorded children, 103 are L1, 64 are L2 speakers.

Along with the speech data, we collected additional metadata with the help of a parent questionnaire. Information such as the children’s first language, their contact time to German, and their age can be useful for speech fluency analyses with respect to different language backgrounds.

3.3 Annotation

The manual annotations are done with Praat [22] within its TextGrid feature on six tiers. As there are many different approaches to annotating disfluencies, depending on field of research and individual research question [23] and no standardised scheme exists, we developed our own

annotation scheme tailored to the specific structure of our speech data resulting from our elicitation method. On the first tier, the orthographic transcript is displayed that has been automatically aligned by Web-MAUS [24] and mainly serves as means of orientation for the annotators. The sixth tier offers space to provide any comments in case the annotators face unusual, difficult to judge or other noteworthy cases to be analysed in the future. Tiers 2 - 5 are used for the annotation of disfluencies.

Tier 2 features the annotation of pauses and interpause intervals. We distinguish between 6 types of pauses, depending on their effect on perceived fluency and their position in the recorded segment. Labels `p_f` and `p_d` are limited to pauses within an utterance. `p_f` represents pauses that were not perceived to interrupt the flow of speech, `p_d` pauses that were perceived to interrupt the flow of speech. Possible factors might be pause length or the position within the utterance. The remaining 4 labels are reserved for pauses between utterances in the recorded segment. `p_start` and `p_end` mark pauses at the beginning and end of the segment to account for turn taking between the virtual character Wuschel's prompt and the child's answer. `p_s` and `p_e` are used to mark pauses after the child's utterances (`p_e`) or before the child's utterances, after an utterance of an external speaker present in the recording room (`p_s`). Intervals between pauses are either labelled as child speech (`sp_ch`) or speech by an external speaker (`sp_ex`).

Tier 3 is reserved for filler particle (FP) annotations. Based on their segmental phonetic structure, we consider 4 types of FPs, consisting of: a) a monophthong vowel ("äh"), b) a monophthong vowel and a nasal ("ähm"), c) a nasal ("hm"), and d) a diphthong vowel ("ei", commonly found in the local dialects)

Tier 4 and 5 are used to annotate a selection of disfluencies, other than pauses and FPs. We decided to label revisions (RV, speech errors which are corrected shortly afterwards), restarts (RS, speech errors which cause the child to abandon their utterance and start over with a new structure), lengthenings (LNGTH, prolongations of sounds, syllables or words), and repetitions (RP, sounds, syllables, or words being repeated in succession, can be interrupted by pauses). Tier 5 covers possible nestings that might occur with these types of disfluencies (e.g. a lengthening within a repetition).

At the time of writing, full annotations were available for 14 (7 L1 German, 7 L2 German) out of the 167 recorded children. This subset was used for training and validation of the model.

4 The Model

4.1 Preprocessing of the Data

While the original manual annotation is highly fine-grained—distinguishing, for instance, multiple subtypes of pauses—we consolidate all pause-related phenomena into a single category for practical modelling. The resulting label set comprises: FP (filler particles), PW (partial words), RP (repetitions), RS (restarts), RV (revisions), P (generic pauses, including all varieties of pauses mentioned in section 3.3), and FLU (fluent tokens). During preprocessing, we map heterogeneous annotation tags to this unified scheme, discard unsupported or ambiguous labels, and segment tokens into utterances. Tokens marked as incomprehensible by human annotators are excluded from the dataset entirely. As shown in Figure 1, the resulting distribution is notably imbalanced, with fluent tokens (FLU) accounting for over half of all samples—a reflection of natural speech patterns. Although such imbalance poses challenges for model training, the substantial size of our dataset (598 training utterances / 5,695 tokens; 150 validation utterances / 1,509 tokens) enables robust learning, as evidenced by the strong performance reported in the following section.

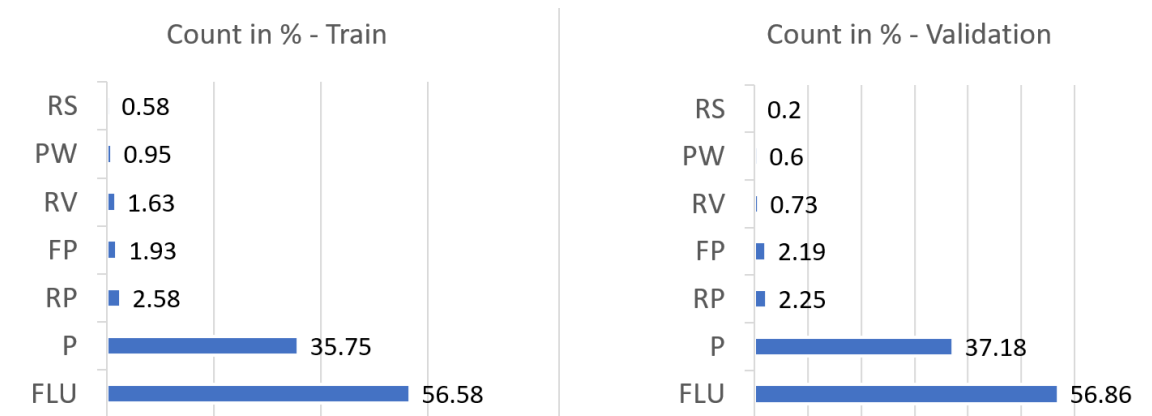


Figure 1 – Counts of each label in Training- and Validation-Dataset: RS = restarts, PW = partial words, RV = revisions, FP = filler particles, RP = repetitions, P = pauses, FLU = fluent tokens

4.2 The Training Process

We frame disfluency detection as a token-level sequence labelling task and fine-tune a pre-trained German BERT model (dbmdz/bert-base-german-cased [25]) for token classification. To prevent data leakage, we perform a sequence-aware train-validation split of the data: 80% of unique segments are used for training and 20% for validation. Tokens are then aligned with BERT’s subword tokenisation by assigning labels only to the first subtoken of each word and masking the rest so they are ignored during loss computation.

5 Results and Discussion

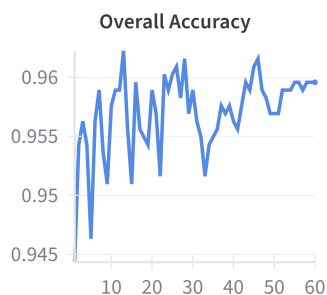


Figure 2 – Accuracy over epochs

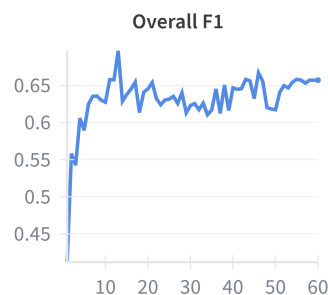


Figure 3 – F1-score over epochs



Figure 4 – Training loss over epochs

Our best-performing model was trained for 60 epochs with a batch size of 16 and a learning rate of 3×10^{-5} , selected based on peak validation macro F1 score. This configuration yields a final token-level accuracy of 96.0% and a macro F1 score of 65.7% on the validation set. For reference, the majority-class baseline (obtained by always predicting the most frequent label ("FLU" - fluent)) achieves an accuracy of 56.86%, confirming that our model substantially outperforms a trivial predictor.

Training dynamics (see Figures 2, 3, and 4) show that loss steadily decreases while accuracy and macro F1 stabilise after approximately 50 epochs, indicating convergence. The relatively modest macro F1 despite high accuracy is primarily due to class imbalance across the seven labels.

As shown in the confusion matrix (Figure 5), performance is strongest for frequent classes (e.g., fluent tokens FLU, pauses P and filler particles FP), while two of the three rarest disfluency

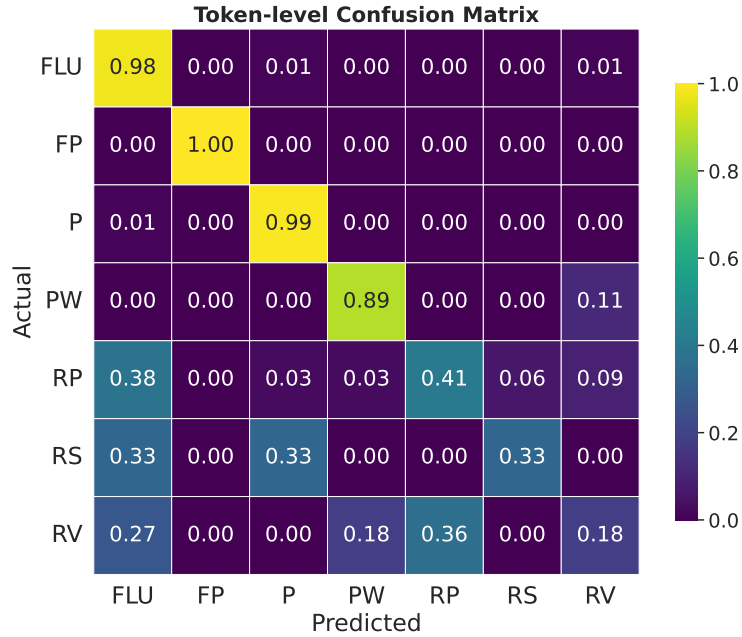


Figure 5 – Normalized confusion matrix for the seven-class disfluency labeling task on the validation set

types, RS and RV, are frequently misclassified as fluent (FLU). This reflects a common bias in imbalanced sequence labeling tasks, where minority classes are collapsed into the dominant class. RVs are most commonly misclassified as RPs. This might be due to the fact that RVs are, in contrast to RSs, often minor corrections, such as function words and inflectional morphemes (e.g. "Du musst die - den Hut holen."). In such cases, the repair frequently consists of a single word whose linguistic properties are similar to those of the reparandum, making confusion with RPs more likely for the model. RSs, on the other hand, are often misclassified as Ps. Both classes tend to occur at utterance boundaries: RSs as utterances are abandoned and restarted with a different structure, and Ps before the start and after the completion of utterances. This overlap in contextual position may influence the model’s predictions.

6 Summary and Future Work

This study was primarily exploratory in nature, aimed at assessing the feasibility of applying transformer-based sequence labelling to disfluency detection in our domain—namely, German child speech—despite significant class imbalance. Encouragingly, even with a coarse label set, our best model achieves a token-level accuracy of 96.0% and a macro F1 score of 65.7%, substantially outperforming the majority-class baseline (56.86%). Error analysis reveals that misclassifications are concentrated almost exclusively on the rarest disfluency types (e.g. RS, RV), which are frequently collapsed into the fluent (FLU) class due to insufficient training examples.

For future work, we plan two key extensions. First, the annotation process of the data is ongoing. As more child speech data is annotated, we will reintroduce finer-grained disfluency categories currently merged or discarded. Preliminary experiments suggest that expanding the label set before sufficient examples per class are available leads to degraded performance, particularly in macro-averaged metrics.

Second, we aim to incorporate an acoustic model (based on WavLM [21]) to capture prosodic and phonetic cues such as vowel lengthening, pauses, and intonation contours. Such features are essential for modelling phenomena that cannot be reliably inferred from text alone (e.g., silent hesitations or elongated syllables), and their integration into a text-signal combined

framework is expected to further improve robustness, especially for underrepresented disfluency types.

Acknowledgements:

We thank Julia Schu and Diana Davidson for preparing the data, and Bernd Möbius and Jürgen Trouvain for their insightful comments on this paper.

References

- [1] ROMANA, A., K. KOISHIDA, and E. M. PROVOST: *Automatic disfluency detection from untranscribed speech*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, pp. 4727–4740, 2024. doi:10.1109/TASLP.2024.3485465.
- [2] RÖHNER, C. and M. WIEDENMANN: *Kinder stärken in Sprache(n) und Kommunikation*. Kohlhammer Verlag, 2017.
- [3] AUTOR*INNENGRUPPE BILDUNGSBERICHTERSTATTUNG: *Bildung in Deutschland 2024: Ein indikatorengestützter Bericht mit einer Analyse zu beruflicher Bildung*. wbv Media GmbH & Company KG, 2024.
- [4] SCHULZ, P. and R. TRACY: *Linguistische Sprachstandserhebung - Deutsch als Zweitsprache (LiSe-DaZ): Language Test for Children with German as a Second Language*. 2011.
- [5] GAGARINA, N., D. KLOP, S. KUNNARI, K. TANTELE, T. VÄLIMAA, U. BOHNACKER, and J. WALTERS: *Main: Multilingual assessment instrument for narratives – revised*. *ZAS Papers in Linguistics*, 63, p. 20, 2019. doi:10.21248/zaspil.63.2019.516.
- [6] MAYR, T. and M. ULICH: *Sismik–Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen. Ein Instrument zur systematischen Beobachtung der Sprachentwicklung*. Freiburg, 2003.
- [7] GINTHER, A., S. DIMOVA, and R. YANG: *Conceptual and empirical relationships between temporal measures of fluency and oral english proficiency with implications for automated scoring*. *Language Testing*, 27(3), pp. 379–399, 2010. doi:10.1177/0265532210364407. URL <https://doi.org/10.1177/0265532210364407>.
- [8] IWASHITA, N., A. BROWN, T. MCNAMARA, and S. O’HAGAN: *Assessed levels of second language speaking proficiency: How distinct?* *Applied Linguistics*, 29(1), pp. 24–49, 2008. doi:10.1093/applin/amm017. URL <https://doi.org/10.1093/applin/amm017>. <https://academic.oup.com/applij/article-pdf/29/1/24/719484/amm017.pdf>.
- [9] REVESZ, A., M. EKIERT, and E. TORGERSEN: *The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance*. *Applied Linguistics*, 37, pp. 828–848, 2016. doi:10.1093/applin/amu069.
- [10] ROCHE, J., S. HABERZETTL, G. PAGONIS, M. JESSEN, and N. WEIDINGER: *Serious Games in der Sprachstandsermittlung*, pp. 340–358. Narr Francke Attempto Verlag, 2019. doi:<http://dx.doi.org/10.22028/D291-35846>.
- [11] KANY, V. and J. TROUVAIN: *Annotation of disfluencies in child speech*. In S. GRAWUNDER (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2025*, pp. 247–254. TUDpress, Dresden, 2025. URL https://www.essv.de/pdf/2025_247_254.pdf.
- [12] VENKATASUBRAMANIAM, L., V. SUNDER, and E. FOSLER-LUSSIER: *End-to-end word-level disfluency detection and classification in children’s reading assessment*. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. 2023. doi:10.1109/ICASSP49357.2023.10095555.
- [13] TRAN, T., M. TINKLER, G. YEUNG, A. ALWAN, and M. OSTENDORF: *Analysis of disfluency in children’s speech*. 2020. doi:10.48550/ARXIV.2010.04293. URL <https://arxiv.org/abs/2010.04293>.

- [14] PROENÇA, J., D. CELORICO, S. CANDEIAS, C. LOPES, and F. PERDIGÃO: *Children’s reading aloud performance: A database and automatic detection of disfluencies*. In *Proceedings of Interspeech 2015*, pp. 1655–1659. ISCA, 2015. doi:10.21437/Interspeech.2015-382. URL <http://dx.doi.org/10.21437/Interspeech.2015-382>.
- [15] LIU, J., A. WUMAIER, D. WEI, and S. GUO: *Automatic speech disfluency detection using wav2vec2.0 for different languages with variable lengths*. *Applied Sciences*, 13(13), p. 7579, 2023. doi:10.3390/app13137579. URL <http://dx.doi.org/10.3390/app13137579>.
- [16] YILDIRIM, S. and S. NARAYANAN: *Automatic detection of disfluency boundaries in spontaneous speech of children using audio–visual information*. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1), pp. 2–12, 2009. doi:10.1109/TASL.2008.2006728.
- [17] AN, G., D. G. BRIZAN, and A. ROSENBERG: *Detecting laughter and filled pauses using syllable-based features*. In *Interspeech 2013*, p. 178–181. ISCA, 2013. doi:10.21437/interspeech.2013-62.
- [18] CHATZIAGAPI, A., D. SGOUROPOULOS, C. KAROUZOS, T. MELISTAS, T. GIANNAKOPOULOS, A. KATSAMANIS, and S. NARAYANAN: *Audio and ASR-based filled pause detection*. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7. 2022. doi:10.1109/ACII55700.2022.9953889.
- [19] GODFREY, JOHN J. and HOLLIMAN, EDWARD: *Switchboard-1 release 2*. 1993. doi:10.35111/SW3H-RW02. URL <https://catalog.ldc.upenn.edu/LDC97S62>.
- [20] DEVLIN, J., M. CHANG, K. LEE, and K. TOUTANOVA: *BERT: pre-training of deep bidirectional transformers for language understanding*. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>. 1810.04805.
- [21] CHEN, S., C. WANG, Z. CHEN, Y. WU, S. LIU, Z. CHEN, J. LI, N. KANDA, T. YOSHIOKA, X. XIAO, J. WU, L. ZHOU, S. REN, Y. QIAN, Y. QIAN, J. WU, M. ZENG, X. YU, and F. WEI: *Wavlm: Large-scale self-supervised pre-training for full stack speech processing*. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), p. 1505–1518, 2022. doi:10.1109/jstsp.2022.3188113. URL <http://dx.doi.org/10.1109/JSTSP.2022.3188113>.
- [22] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer (version 6.4.04)*. 2024. URL <http://www.praat.org>.
- [23] TROUVAIN, J., L. CRIBLE, M. BELZ, S. BETZ, ŠTEFAN BEŇUŠ, L. BAQUÉ, M. CANTARUTTI, J. D. NAPOLI, I. DIDIRKOVÁ, M. MACHUCA, L. MAREKOVÁ, O. NICULESCU, P. PELTONEN, A. PISTONO, L. SCHETTINO, V. SILBER-VAROD, and S. WILLIAMS: *On variability in the identification and labelling of disfluencies — preliminary results from 23 annotations of the same data*. In *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*, pp. 57–61. Lisbon, 2025. doi:10.21437/DiSS.2025-12.
- [24] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. 14th International Congress of Phonetic Sciences (ICPhS)*, pp. 607–610. San Francisco, 1999.
- [25] BAYERISCHE STAATSBIBLIOTHEK: *bert-base-german-cased (revision 43ccea13)*. 2025. doi:10.57967/hf/4377. URL <https://huggingface.co/dbmdz/bert-base-german-cased>.