Interspeech 2022
18-22 September 2022, Incheon, Korea

kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech

Lars Rumberg[1], *Christopher Gebauer*[1], *Hanna Ehlert*[2], *Maren Wallbaum*[2], *Lena Bornholt*[2], *Jörn Ostermann*[1], *Ulrike Lüdtke*[2]

[1]Institut für Informationsverarbeitung - L3S, Leibniz University Hannover, Germany
[2]Institut für Sonderpädagogik, Leibniz University Hannover, Germany

kidstalc@tnt.uni-hannover.de

Abstract

In this paper we present kidsTALC an audio dataset with orthographic and phonetic transcriptions of German children's speech collected to facilitate the development of speech based technological solutions. The dataset is part of a larger project aiming to develop machine-learning applications to support automation in child speech and language assessment for research and clinical purposes. At the same time, the interdisciplinary project was established to increase the accessibility of corpora of continuous child speech in Germany and globally to train accurate automated speech recognition tools for children. In the first stage we collected and transcribed 25 hours of continuous speech from typically developing children aged 3 ½–11 years. Here, we discuss the key features of the dataset, data collection, transcription protocol and future datasets in the project. We also present important statistics of our dataset and will demonstrate the speech recognition performance of one baseline model on the dataset.

Index Terms: speech corpus, child speech, German speech, speech recognition

1. Introduction

In recent years, the performance of automatic speech recognition (ASR) systems drastically increased and allows the usage in a wide variety of everyday applications. Even though these advances are promising, similar results for children's speech have not been achieved yet. Multiple concerns lead to this decrease in performance. One reason is the higher variability between and within child speakers, compared to adult speakers, due to their smaller anatomic structures of the vocal tract, developing motor control as well as phonological proficiency [1]. Additional to this intrinsic difficulty of children's speech, a major problem is that large scale children's speech datasets are required to train and test end-to-end applications [2], even when incorporating out of domain adult speech [3, 4]. While the performance of ASR on children's speech is lacking behind, the necessity for such systems is even greater. Especially, the opportunities for automated support in child speech and language assessment are very promising [5].

Datasets for children's speech that are publicly available are very scarce, especially when specific languages and/or connected natural speech are targeted. The PF_STAR Children's speech corpus [6] contains mainly read speech from English, German, Italian and Swedish speaking children. The corpus includes a small subset of free speech, but it's recordings of German children are limited to children aged 10 and above. Furthermore, the corpus does not provide a phonetic transcription. The OGI Kids' speech corpus [7] is supposed to target all common American English biphones, by recording a variety of children being well distributed across age and gender. The dataset is mainly limited to scripted speech with a limited set of unique sentences and isolated words, but contains a small part of unscripted speech as well. In the National Institute of Technology Karnataka Kids' (NITK Kids') Speech Corpus [8] recordings of 160 children on a picture description task, to retrieve specific isolated words, are present. However, the data is limited to the Kannada language and very young children (max. 6 ½). The childLex corpus [9] contains German children's read speech and is of large scale, but lacks any manual transcriptions.

In more recent years, smaller datasets appeared, focusing on free speech combined with manual transcriptions. However, the scarcity is obvious as the cost for manual transcriptions is high [10]. The TLT-school corpus [11] contains speech from Italian children with the ages 9 and above, participating in German or English class as a second language. The recordings are part of a second language proficiency test, where all pupils are recorded at once in their class rooms, which adds a high percentage of background noise and overlapping speech. Unfortunately, the manual transcriptions are limited to orthographic. The AusKidTalk corpus [12] includes recordings from Australian children ageing from 3–12 years. The recording setting is restricted to child-computer-interactions covering isolated words, story telling and question answering, but lacks any natural communication in interaction. Orthographic transcriptions are available for all elicitation types, however only the scripted speech is annotated on phone level. The corpus does not include German speaking children. For a further listing of available datasets, we refer to Ramteke *et al.* [8].

Another important source for speech corpora is TalkBank, where the CHILDES project [13] is the part focusing on children's speech. Within this project multiple smaller datasets are made (mostly) publicly available, however many of those only provide transcriptions and no audio. The part containing phonetic transcriptions coupled with audio is located in the PhonBank [14]. For the German speaking children in the PhonBank the ages are either restricted to very young children (max. 4 ½) or limited to picture naming tasks. More free speech is found in the HomeBank [15], but it does not contain German speech and has very restricted access.

To overcome some of these limitations for publicly available datasets for children's speech, we present kidsTALC. The repository is originated in the project *Tool for Analyzing Language and Communication (TALC)*. The specific focus of the first dataset of our kidsTALC repository is typically developing, monolingual German children. As we aim to develop machine-learning applications to support automation in child speech and language assessment for research and clinical pur-

10.21437/Interspeech.2022-330

poses, we have major requirements towards our own dataset. This results in the following contributions, which emphasize the necessity of our dataset:

- A collection of connected, German speech from children of various age and language status
- Manual, revised, orthographic and phonetic transcriptions of all the collected speech data
- Annotations of the scripted speech for developmental errors regarding phonetics, phonology, syntax, morphology and semantics
- Public access to the kidsTALC repository for research purposes[1]

2. Data Acquisition

The motivation to collect kidsTALC is the training of machine-learning applications to support language sample analysis for research and clinical purposes, with the elicitation contexts focus on natural, spontaneous child speech. Therefore, a number of requirements arise towards our dataset. As the transfer from ASR trained on isolated speech is difficult, the recording setup needs to focus on connected natural speech and cover typical clinical elicitation contexts used in speech and language assessment, as well as everyday speech production of children. Additionally, a large scale of data samples and a wide range of different speakers is necessary to allow robust training of machine-learning applications. All data needs to be transcribed orthographically and phonetically, to allow a variety of analysis, e. g., regarding speech sound error patterns or incorrect syntax. To increase the resulting performance of the machine-learning application the quality of the transcriptions needs to be as high as possible, especially concerning the inter-transcriber variance. The repository needs to provide annotations regarding all necessary error patterns, e. g., on phonetics, syntax or morphology level, with enough speaker examples. In the following section, we will describe how we addressed all the requirements, while keeping balance between high quality of the data and annotation costs.

2.1. Participants

Participants for the entire repository are being recruited from a network of collaborating preschools, kindergartens and elementary schools. Prior to participation, caregivers provide a written consent for inclusion in the kidsTALC repository. Verification of eligibility criteria of the participating children is obtained from caregivers and teachers. Eligibility criteria for the first dataset presented here are: 3 ½–11 years, monolingual German speakers, typically developing. Additionally, children are excluded from the dataset, if the examiners have concerns about their language development during data collection. However, the dataset includes children with age-related developmental speech and language errors typical for monolingual acquisition of German on all linguistic levels, such as phonological errors, case marking errors or neologisms [16, 17]. The dataset is stratified for these errors to include at least four examples of all typical errors in this age range. The dataset is divided in four age groups: 3;6–4;11 (AG1), 5;0–6;11 (AG2), 7;0–8;11 (AG3), 9;0–10;11 (AG4) years. The notation for the age is in the format *years;months*. The duration of the recordings, including speech of the child and the examiner, ranges from 30–60 min, whereas the proportion of child utterances is on average 15.6 min (ranging in 5–30 min). A more detailed distribution is found in Sec. 3.2.

[1] To apply for the dataset, please contact us or visit `https://www.tnt.uni-hannover.de/en/project/talc/`.

2.2. Setting

The samples are collected in an examiner-child interaction. Examiners are trained speech language therapists or speech language therapy students. The recordings take place at home or in the kindergarten/preschool of the children in a quiet room with no or little background noise. Few recordings include utterances of a second child (e. g., a sibling), but are marked as such.

Material for sample elicitation are seven different wordless picture books appropriate for children of the different age groups, e. g., „Quest“ [18], „All Around Bustletown: Spring“ [19], “Good Night Gorilla” [20]. In some cases the children additionally bring own books to the recording session. The children have free choice of the books they want to look at with the examiner or engage in conversational discourse on topics of their choice spawned by the books' content. The elicitation context therefore varies between narrative (story telling), picture description and conversational discourse. Tied to these different elicitation contexts is a various degree of spontaneity of the connected speech in the samples. Story telling and picture description represent to a lesser degree natural communication, because they are language tasks, whereas the parts with conversational sampling reflect intrinsically motivated communicative interaction. This impacts the speech itself, for example fluency, pitch or stress and can thereby influence automated processing.

The examiner protocol for data collection includes two main aspects. Firstly, to ask open-ended questions, especially external state questions (e. g., “What's happening?”), as they are thought to facilitate (complex) speech production [21], and secondly, to avoid overlapping speech, maximizing the amount of processable child utterances. For audio recording an Olympia dictaphone (LS-P1) is utilized and placed close to child's mouth (distance of 10–50 cm). The sound is stored in an uncompressed format (.WAV) at a sampling frequency of 96 kHz and a bit depth of 24 bit.

2.3. Data Transcription and Annotation

Manual transcriptions and annotations of all recordings are compiled. Adult and child utterances are transcribed orthographically (standard German) and additionally child utterances are transcribed phonetically (verbatim). Overlapping, unintelligible and non-verbal (e. g., laughing or coughing) parts are marked. Systematic developmental errors regarding speech sound, grammar and vocabulary, as well as elicitation context are annotated.

Transcriptions and annotations are completed by trained graduate students of speech language therapy, trained speech language therapists and a professional transcription agency. To ensure reliability and consistency of all transcriptions and annotations, a transcription protocol incorporating a three step procedure for each recording is applied. Initial orthographic and phonetic transcription are generated and, secondly, checked by a different transcriber. Finally, the phonetic transcriptions are checked by a professional transcription agency. In case of transcription differences, the concerning passage is discussed until consensus is reached.

In addition, to account for the developmental errors that are non-perceptible purely based on the recordings, all audios are accompanied by metadata. These metadata contain informa-

tion about voice quality (e. g., nasality due to infections of the respiratory tract) and speech sound errors identified by the examiners while conducting the data collection. Developmental speech and language errors are annotated manually and checked by a second annotator. Additional annotation regarding elicitation context (in this dataset: picture description, story telling, conversational discourse) is conducted once, if the dataset contains more than one elicitation context. In a subsequent step all recordings are anonymized to protect the identity of the children, i. e., all names or personal information are replaced with a special token in the transcriptions and the audio is corrupted with silence. For transcription and annotation the ELAN tool [22] is utilized.

2.4. Phonetic Token Set

The project uses a simplified chart of the International Phonetic Alphabet (IPA) for German pronunciation, to balance required detail and practicability of the phonetic transcriptions. The phone selection is guided by a dual perspective. On the one hand, from a speech language therapist perspective, the set should be sufficient enough to allow the identification of speech sound errors in typically developing children and in children with speech, language and communication needs, such as developmental language disorders or speech sound disorders. On the other hand, IPA elements not seen relevant for screening purposes could be omitted, reducing the difficulties that arise for the ASR software, as well as the difference between two transcribers. Therefore, the project's IPA set aims for a broad transcription [23]. It excludes diacritics, suprasegmentals except vowel length marking, and some IPA characters for German, such as semivowls and few consonants, e. g., the glottal stop /ʔ/.

3. Statistical Analysis

In this section we summarize key features of the first dataset in our kidsTALC repository. As mentioned before, the quality of the transcriptions is important, especially if robust software based on machine-learning is targeted. To quantify the variance in the transcriptions without any control steps, we will perform a transcriber agreement study. This will demonstrate errors, which humans have problems separating and models will most likely have as well. Closing, we will show the age, word, and phone distribution, respectively.

3.1. Transcriber Agreement

To analyse the agreement of the transcribers for the phonetic transcription, we randomly selected six short, six average, and six long utterances for both male and female speakers of each of the four age groups. These 144 utterances were then transcribed by three of the transcribers. We computed the phone error rate (PER) between each pair of the three transcriptions, as well as the PER to the final transcript, which has undergone all steps described in Sec. 2.3. The average inter-transcriber PER is $14.6\,\%$. The average PER of a transcript to the final transcript is $12.8\,\%$. The most common disagreements are related to the elongation mark on vowels, substitutions of similar vowels (/e/ and /ə/ with /ɛ/ and /a/ with /ɐ/), substitutions of the nasals /n/ and /ŋ/; and deletions/insertions of trailing /t/ sounds especially for the German words *und* and *jetzt*, for which the final /t/ is also commonly omitted in adult speech. These disagreements have only minor relevance for child speech and language assessment. Furthermore, we expect the final transcript to be more consistent than these numbers suggest due to our multi-level transcription process. Nevertheless, the user of the corpus should remain aware of these imperfections of the phonetic transcription.

3.2. Age Distribution

In Tab. 1 the speakers from the first dataset are displayed. Of special interest are the younger children of course, due to higher variance in speech production. For the test set a boy and a girl of each age group are selected. We also suggest a development subset of the train set with a similar distribution as the test set. The ground truth for the test set is not publicly available, but a leader board will be hosted on our repository's web page.

Table 1: *Age distribution separated by our four age groups and sex. We display the total quantity of children's speech, as well as the subset of the corpus ignoring utterances with hard to understand, or overlapping speech in round brackets, both in minutes. The number of speakers is in square brackets.*

age	sex	train	test
3;6 - 4;11	f	84.1 (62.7) [6]	17.2 (14.9) [1]
	m	67.5 (45.6) [4]	17.8 (10.7) [1]
5;0 - 6;11	f	153.8 (98.9) [8]	22.2 (10.2) [1]
	m	149.4 (92.6) [8]	16.7 (12.7) [1]
7;0 - 8;11	f	22.5 (12.8) [2]	14.1 (7.5) [1]
	m	46.0 (26.2) [3]	12.3 (11.2) [1]
9;0 - 10;11	f	30.6 (22.2) [3]	9.7 (8.6) [1]
	m	74.6 (44.8) [5]	11.3 (8.2) [1]

3.3. Word Distribution

In the first part of the kidsTALC dataset ~ 4300 unique words with ~ 7600 unique pronunciations are used over a total of around 55 thousand uttered words. This represents the proportions in round brackets of Tab. 2. Fig. 1 demonstrates the word and pronunciation distribution, especially that the dataset only contains few samples for most of the pronunciation variants.

3.4. Phone Distribution

Fig. 2 demonstrates the phone distribution in the first part of our kidsTALC repository. In total 176 thousand phones are uttered, distributed over 40 different phones. These numbers represent the subset of the corpus in round brackets of Tab. 2.

4. Automatic Speech Recognition

In this section we demonstrate the usage of our dataset on an ASR model trained end-to-end. We implemented the baseline using SpeechBrain [24]. The feature extractor computes Mel spectrograms based on the raw audio, with a window size of 25 ms and a hopping length of 10 ms. The spectrograms are processed by multiple convolutional layers and then in turn passed to a bi-directional recurrent neural network. Finally, the output is processed by a dense layer to compute the softmax function over our phone set for each time frame. The model is trained using the connectionist temporal classification (CTC) loss [25]. In Tab. 2 the PER is shown for our train and test set.

To ease reproduceability, we used for training the recipe from SpeechBrain [24] found at *recipes/TIMIT/ASR/CTC*. From this recipe, we adjusted the dataset pipeline, to fit our dataset. Furthermore, we adjusted learning rate ($l_r = 0.0003$), optimizer (Adam [29]), and the scheduler (OneCycleLR [30]). To

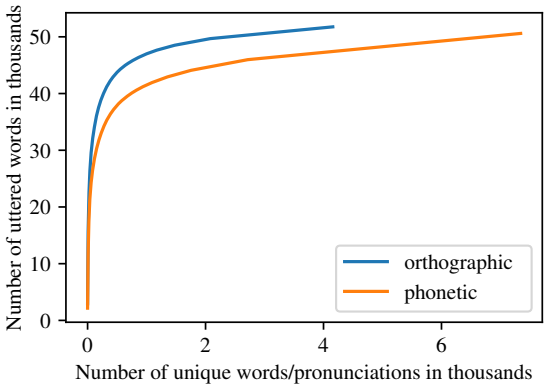

Figure 1: *Cumulative graph of total uttered words over unique words or pronunciations, sorted by frequency. Orange represents the phonetic transcription, where a few unique pronunciation variants account for a high fraction of the total uttered words. The flat end of the graph demonstrates that the dataset contains many words with only few samples. The blue line represents the orthographic transcript, the major difference is that multiple pronunciations exists for a single orthographic word.*

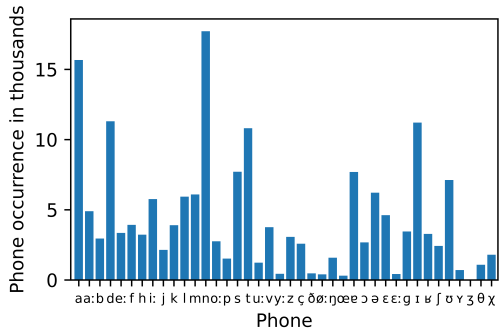

Figure 2: *Occurrence of each phone in the first dataset of the kidsTALC repository. Not all phones, especially the ones representing mispronunciations, e. g., lisping, are present, as the datasets contains only typically developing children.*

further stabilize the training we removed all the augmentation and only applied frequency masking [31]. However, the optimization algorithm stays unchanged and we have not tuned any of the other hyperparameters, e. g., the model structure.

The high values of the PER demonstrate the difficulties that arise with children's speech. The lower PER of the test split compared to the dev split can be explained by difficulties for single children. The PER per speaker ranges from under 15 % to over 40 %. Including adult speech, and therefore increasing the size and diversity of the train set, does have a positive effect.

5. Recording Status and Future Datasets

The dataset presented here is only a first step towards a greater repository, which will be extended in the next years to target automation in child speech and language assessment. By now we are targeting monolingual German speakers, without a regional dialect, which are typically developing. In the future we will include children with spoken and written language disorders, younger children, children with other first languages, e. g., Afrikaans, and various elicitation contexts of connected speech. A short summary of the already timed datasets, or datasets which we are already recording, is given in Tab. 3.

Table 2: *Phone error rate on our development and test set for a baseline model trained using SpeechBrain [24]. We trained the model purely based on our dataset and in combination with the mozilla common voice (MCV) dataset [26]. We translated the orthographic transcriptions in MCV using an external pronunciation dictionary [27] based on data from BAS [28] to create phonetic labels.*

	Dev	Test
kidsTALC	35.75	26.18
kidsTALC + MCV	32.50	24.04

Table 3: *Recording status of all planned datasets, part of our kidsTALC repository. The presented dataset is in the first row and will be extended in the near future. Besides the estimated year of completion, the final number of kids and the number which have been fully transcribed are stated. The recording type, i. e., spontaneous (S) or read (R) speech, as well as the language status of the included children, differing between typically developing (TD), developmental language disorder (DLD), speech sound disorder (SSD) and reading difficulty (RD) is given. The age range of the participating children is stated last.*

Date	Tot.	Comp.	Type	Dev.	Age
2022	90	47	S	TD	3;6–10;11
2023	40	0	S	TD	3;0–7;0
2024	60	0	S	DLD/SSD	3;0–7;0
2024	100	0	R	TD/RD	8;0–10;0

6. Conclusions

kidsTALC is the first German speech corpus that addresses the modern standards to meet the requirements for developing automatic tools to support language sample analysis in research and clinical applications. The repository consists of multiple datasets (all containing connected speech), to represent different recording settings, language status, and ages. In the final version the repository will contain recordings from about 300 children (of which 47 are finished), while their age range will span Kindergarten to elementary school. The elicitation contexts will cover various settings along the unstructured-structured continuum, such as free play, story tell, conversational discourse or read texts with a focus on spontaneous language. Also children with various oral and written language abilities will be included in the corpus, such as typically developing children and children with developmental language disorder or speech sound disorder. kidsTALC promises to have great impact on the development of machine-learning applications to support automation in child speech and language assessment in German speaking regions.

7. Acknowledgements

We would like to thank all families and children who participated in creating this dataset supporting the kidsTALC repository.

8. References

[1] M. E. Beckman, A. R. Plummer, B. Munson, and P. F. Reidy, "Methods for eliciting, annotating, and analyzing databases for child speech development," *Computer Speech & Language*, vol. 45, pp. 278–299, 2017.

[2] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proceedings INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*. ISCA, 2015, pp. 1611–1615.

[3] D. Smith, A. Sneddon, L. Ward, A. Duenser, J. Freyne, D. Silvera-Tawil, and A. Morgan, "Improving Child Speech Disorder Assessment by Incorporating Out-of-Domain Adult Speech," in *Proceedings INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*. ISCA, 2017, pp. 2690–2694.

[4] L. Rumberg, H. Ehlert, U. Lüdtke, and J. Ostermann, "Age-Invariant Training for End-to-End Child Speech Recognition Using Adversarial Multi-Task Learning," in *Proceedings INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*. ISCA, 2021, pp. 3850–3854.

[5] M. Shahin, U. Zafar, and B. Ahmed, "The Automatic Detection of Speech Disorders in Children: Challenges, Opportunities, and Preliminary Results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2020.

[6] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR Children's Speech Corpus," in *Proceedings INTERSPEECH 2005 – 6th Annual Conference of the International Speech Communication Association*. ISCA, 2005.

[7] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 564–567.

[8] P. B. Ramteke, S. Supanekar, P. Hegde, H. Nelson, V. Aithal, and S. G. Koolagudi, "NITK Kids' Speech Corpus," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*. ISCA, 2019, pp. 331–335.

[9] S. Schroeder, K.-M. Würzner, J. Heister, A. Geyken, and R. Kliegl, "childLex: A lexical database of German read by children," *Behavioral Research*, vol. 47, pp. 1085–1094, 2015.

[10] S. L. Pavelko, R. E. Owens, M. Ireland, and V. D. L. Hahs, "Use of Language Sample Analysis by School-Based SLPs: Results of a Nationwide Survey," *Language, Speech, and Hearing Services in Schools*, vol. 47, no. 3, pp. 246–258, 2016.

[11] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "TLT-school: A Corpus of Non Native Children Speech," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 378–385.

[12] B. Ahmed, K. J. Ballard, D. Burnham, T. Sirojan, H. Mehmood, D. Estival, E. Baker, F. Cox, J. Arciuli, T. Benders, K. Demuth, B. Kelly, C. Diskin-Holdaway, M. Shahin, V. Sethu, J. Epps, C. B. Lee, and E. Ambikairajah, "AusKidTalk: An Auditory-Visual Corpus of 3- to 12-year-old Australian Children's Speech," in *Proceedings INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*. ISCA, 2021, pp. 3680–3684.

[13] B. Macwhinney, "The CHILDES project: Tools for analyzing talk," *Child Language Teaching and Therapy*, vol. 8, 2000.

[14] Y. Rose and B. Macwhinney, "The PhonBank Project: Data and Software-Assisted Methods for the Study of Phonology and Phonological Development," in *The Oxford Handbook of Corpus Phonology*, 2014, pp. 380–401.

[15] M. VanDam, A. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, and B. MacWhinney, "HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings," *Seminars in Speech and Language*, vol. 37, no. 02, pp. 128–142, 2016.

[16] A. V. Fox-Boyer, "German speech acquisition," in *The International Guide to Speech Acquisition*, S. McLeod, Ed. Thomson Delmar Learning, 2007, ch. 41.

[17] D. Bittner, "Case Before Gender in the Acquisition of German," *Folia Linguistica*, vol. 40, no. 1-2, pp. 115–134, 2006.

[18] A. Becker, *Quest*. Candlewick, 2014.

[19] R. S. Berner, *All Around Bustletown: Spring*. Prestel, 2019.

[20] P. Rathmann, *Good Night, Gorilla*. G.P. Putnam's Sons, 1996.

[21] R. Jean-Baptiste, H. B. Klein, D. Brates, and N. Moses, "What's happening? And other questions obligating complete sentences as responses," *Child Language Teaching and Therapy*, vol. 34, no. 2, pp. 191–202, 2018.

[22] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A Professional Framework for Multimodality Research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), 2006.

[23] J. P. Stemberger and B. M. Bernhard, "Phonetic Transcription for Speech- Language Pathology in the 21st Century," *Folia Phoniatrica et Logopaedica*, vol. 72, pp. 75–83, 2020.

[24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," 2021.

[25] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," *International Conference on Machine Learning (ICML)*, p. 8, 2006.

[26] "Mozilla Common Voice," https://commonvoice.mozilla.org/en, 2022.

[27] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[28] "Bavarian Archive for Speech Signals (BAS)," http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B, 2013.

[29] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of 3rd International Conference for Learning Representations (ICLR 2015)*, 2015.

[30] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*. ISCA, 2019, pp. 2613–2617.