Interspeech 2022
18-22 September 2022, Incheon, Korea

Improving Phonetic Transcriptions of Children's Speech by Pronunciation Modelling with Constrained CTC-Decoding

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

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

{rumberg, gebauer}@tnt.uni-hannover.de

Abstract

Language sample analysis (LSA) is a powerful tool for both therapeutic applications and research of child speech and language development. Nevertheless, it is not routinely used, due to the high cost of manual transcription and analysis. Assistance by automatic speech recognition for children has the potential to enable a wide-spread use of LSA. However, the development of modern speech recognition systems heavily relies on large scale datasets. Therefore, it faces the same obstacle of high cost for transcription as LSA itself. In this paper, we study how cheaply transcribed child speech, i. e., limited to an orthographic transcription, can be improved on a phonetic level by leveraging a CTC based automatic speech recognition model, trained on a small phonetically transcribed dataset. We constrain the CTC decoding by modeling variation of the pronunciation given the orthographic transcription using weighted finite state automata. Our experiments show that the transcription is improved in terms of phone error rate by relative 14 % when applying our method. Additionally, we show how the improved transcript can in turn be leveraged to improve the training of a new model.

Index Terms: speech recognition, constrained decoding, child speech, finite state automaton

1. Introduction

Nowadays, automatic speech recognition (ASR) systems address a wide range of applications due to the recent improvements introduced by the progress in deep learning (DL) technologies. However, for children's speech these performance gains are still missing. A possible reason for these shortcomings is the higher intra- and interspeaker variability for children [1]. One possibility to overcome the difficulties in automatic speech recognition of children is to use large scale datasets [2]. While this would be desirable, the cost of such datasets, especially when manually transcribing and annotating phones and miss-pronunciations, are enormous [3]. The amount of data needed can be reduced by incorporating out of domain data, e. g., adult speech [4, 5], but domain-specific data is still needed in non-negligible quantities.

Recently, we introduced the kidsTALC corpus [6] of German children's speech, designed to train ASR systems to be used to facilitate research of speech and language development and assist therapeutic applications. These applications require an accurate phonetic transcription and rich annotations. During the data collection the work necessary to transcribe an audio phonetically is 10–30 times higher than for an orthographic transcription and needs to be done by an expert. It is therefore desirable to reduce the requirement for data with a manual phonetic transcription as much as possible. In this work we focus on how to improve a phonetic transcriptions generated by low-cost orthographic transcriptions and a pronunciation dictionary. Further, we investigate the effect of this improvement on the training of a phonetic ASR system for child speech.

A related topic is the usage of unlabelled speech data, which has been proven to be effective many times in recent years on the Libri Light benchmark [7]. Popular methods include iterative self-training [8] and unsupervised pre-training [9]. The former trains a series of models in an iterative fashion, where each model generates pseudo-labels on the unlabeled data, which is then used as labels for the succeeding model. The latter trains a huge encoder using a contrastive task on unlabeled data to learn speech representations. This encoder is then fine tuned on the labeled speech. Zhang *et al.* [10] combine self-training with unsupervised pre-training and show that both approaches are complementary.

On the other hand is lightly supervised learning, which utilizes data with low quality labels, an extensively researched topic for adult speech. Examples are Olcoz *et al.* [11] and Fainberg *et al.* [12]. Both aim to improve the quality of poor subtitles in a broadcast dataset. While Olcoz *et al.* improve the alignment using the Viterbi algorithm, Fainberg *et al.* use weighted finite state automata (WFSA) to combine the transcriptions, an edit transducer allowing for specific error patterns, and the lattice of a GMM-HMM model. The combined WFSA is used as supervision during training of the acoustic model.

Improving a pseudo-phonetic transcription can be done by identifying pronunciation deviations and correcting them in the transcript. Multiple approaches to identifying pronunciation errors for children's speech exist in prior work. Yilmaz *et al.* [13] proposed a data-driven approach for child reading assessment. The authors compute a phone confusion matrix on the train set of the data to account for possible errors made during speech production. The substitution probabilities are used as costs in an WFSA to allow deviations from the expected words. This approach is limited to errors present in the training corpus, and cannot deal with unseen error patterns. Instead of learning error patterns from the data Ward *et al.* [14] use expert speech language therapists knowledge by modelling commonly known phonological error patterns (PEPs). Results are presented on a small dataset consisting of a limited number of isolated words. The relevance of PEP is confirmed by Fringi *et al.* [15]. The focus of the authors was to train a baseline model and identify errors with significant relevance. However, the authors noted that only 7–8% of the errors made by the ASR model match with the commonly known PEPs.

Whereas the before mentioned approaches mainly focus on identifying errors during decoding, Nicolao *et al.* [16] proposed an approach for reading assessment that uses the decoding results to also improve the training of the ASR model. The au-

10.21437/Interspeech.2022-332

thors constrain the GMM-HMM based latticed with the read text by utilizing a WFSA. As the optimization depends on the quality of the available data, the authors improved the artificially labeled data in iterative fashion, similar to Xu *et al.* [8]. Chu *et al.* [17] also apply an iterative optimization scheme. The authors identify mispronounced phones with the goodness-of-pronunciation (GOP) measure [18] and correct them in the transcription using an edit-distance transducer.

All of the above methods for pronunciation error detection use hybrid HMM-GMM or HMM-DNN models. We, in contrast, use connectionist temporal classification (CTC) to optimize our end-to-end DL model and combine it with a WFSA based decoding scheme. Decoding CTC based models with WFSA has been shown to be effective by Miao *et al.* [19]. To our best knowledge, we are the first that use this approach for modelling variation of the pronunciation in the decoding graph and apply it to child speech.

2. Constrained Decoding

In the present work, we aim to minimize the required amount of phonetically transcribed data, due to its high cost. Orthographic transcriptions are much cheaper to compile, but are limited to an idealized pronunciation from a dictionary. Especially for young children it must be assumed that the real speech deviates from the standard pronunciation a lot. Directly using these transcriptions as training data therefore leads to an idealized ASR system, which is biased towards the standard pronunciation and poorly captures variations. We aim to recover variation of the speech production by extending the idealized pseudo-phonetic transcriptions by possible error patterns and weight them using the lattice generated from an ASR model, trained using the CTC criterion and a small speech corpus with a manual phonetic transcription. We do so by utilizing weighted finite state automaton (WFSA) to efficiently combine all sources of information. In the following, we assume an orthographic transcription is present for all recordings.

2.1. Decoding Graph

In this section we will describe, how we apply weighted finite state automaton (WFSA) in our system. In general a WFSA is a directed graph used to efficiently represent sequences. For more details towards the structure, mathematical operations, and implementation details we refer to further literature [20, 19].

We use WFSA to represent sequences of phones. Each transition accepts a phone or a blank label and carries a weight, representing the log-likelihood of this transition. The simplest way to decode the output of an acoustic model, trained with the CTC criterion, is greedy decoding. We use WFSA to describe greedy CTC-Decoding as finding the shortest path in the decoding graph $\mathcal{G}$

$$\mathcal{G} = \mathcal{C} \circ \mathcal{H}, \tag{1}$$

where $\mathcal{H}$ is a dense emission graph of the acoustic model, and $\mathcal{C}$ is a graph modeling the CTC-Algorithm. It transduces repetitions and emissions of the blank label to no output. $\mathcal{G}$ accepts all possible sequences given the token-set, limited only by the length of the model output.

To constrain $\mathcal{G}$, we introduce a graph $\mathcal{S}$ and compute its composition with $\mathcal{C}$ before computing the composition with $\mathcal{H}$:

$$\mathcal{D} = (\mathcal{C} \circ \mathcal{S}) \circ \mathcal{H}. \tag{2}$$

By defining $\mathcal{S}$, such that it only accepts some sequences of phones, we constrain the decoding to these sequences. This allows us to use $\mathcal{S}$ to constrain the decoding towards the available orthographic transcription.

To compute $\mathcal{S}$, we represent the pronunciation of a word w as a linear graph $\mathcal{P}_w$ with a transition for each phone in the pronunciation. Given the orthographic transcription of an utterance and a pronunciation dictionary with a single pronunciation variant for each word, we concatenate the pronunciation graphs $\mathcal{P}_w$ of each word in the utterance. Using the resulting graph as $\mathcal{S}$, all paths in the decoding graph $\mathcal{D}$ have the same output and differ only in the CTC-alignment. By extending the graph $\mathcal{S}$, deviations from the standard pronunciation can be modelled.

2.2. Pronunciation Modelling

The described decoding scheme allows multiple pronunciation variants, as the sequence graph $\mathcal{S}$ is not limited to one linear graph. In this section we will describe how we incorporate meaningful alternatives to the standard pronunciation present in the dictionary.

2.2.1. Pronunciation Variants from Data

In the first step, we extend the pronunciation dictionary by collecting all pronunciations in the phonetically transcribed part of the train set, to cope for common pronunciations used by children. This allows us to utilize multiple pronunciations, which have high probability around our target group of speakers. Afterwards, we consider for each word up to N most common pronunciation variants from the extended pronunciation dictionary. We again compute the sequence graph $\mathcal{S}$ by concatenating $\mathcal{P}_w$, which now represents the union of all pronunciation graphs $\mathcal{P}_{w,i}$ with $i \in [1, ..., N]$. If less than N variations are available in the pronunciation dictionary, we consider all for this specific word. The resulting sequence graph accepts all combinations of pronunciations for the words in the utterance.

We include prior knowledge about the likelihood of a pronunciation i of a word w by setting the weight of the first transition in the pronunciation graph $\mathcal{P}_{w,i}$. For all words seen in the train set, we compute the pronunciation's frequency normalized by the words frequency. If the word is not present in the train set, we only use the standard pronunciation of an external dictionary and no weight is necessary. The shortest path in the resulting decoding graph $\mathcal{D}$ represents the sequence of the most likely pronunciation for each word, re-weighted given the output of the acoustic model.

2.2.2. Data Driven Substitutions, Deletions and Insertions

Similarly to Yimlaz *et al.* [13] we insert deviations from the standard pronunciations based on common error patterns in the data. We identify common pronunciation deviations in the train set by computing the Levenshtein alignment between the standard pronunciation from a pronunciation dictionary as reference and the manual phonetic transcription as hypothesis. While the strategy described in Sec. 2.2.1 models pronunciation variations on a word level, the approach described here models them on a phone level. We introduce substitutions, deletions, and insertions into the decoding graph by adding additional transitions to the sequence graph $\mathcal{S}$ allowing the WFSA to accept the modified sequence.

Using the Levenshtein alignment between the standard pronunciation and the manual transcription we compute the weight for each added transition by counting the occurrences of this substitution, deletion or insertion in the train set, e. g., how often a /s/ is replaced by a /z/. For deletions and insertions we

also take the neighboring phones into account. We normalize the weight by the total frequency of the original phone (for substitutions), bi-phones (for insertions), respectively tri-phone (for deletions). Finally, we adjust the weight of the original transitions in $\mathcal{S}$, such that the total score stays unchanged. To avoid an unnecessary huge decoding graph, we only consider modifications with a relative frequency in the train set of above $5\,\%$.

2.2.3. Phonological Error Patterns

Using expert domain knowledge from speech language therapists by modelling PEPs has been shown to be effective [14, 4]. Such deviations follow strict rules [21] and modify the pronunciation of a given word based on the present syllables. Their prevalence is a characteristic of the speaker's age. The syntax of the sentence, especially the position of the spoken word, does not have an influence. As the deviated words will most likely not be present in the pronunciation dictionary, we apply the PEP-function directly to the phonetic transcription. Additionally, we only consider non-pathological error patterns, as our dataset is limited to typically developing children. The relevant non-pathological PEPs for the present age groups, which we model, are fronting, reduction of initial consonant clusters, and sigmatism/lisp. However, this can be easily extended to any pattern, if the dataset contains children with developmental language disorders or speech sound disorders. We add the deviated words to $\mathcal{S}$ in an identical fashion as the pronunciation variants from the pronunciation dictionary described in Sec. 2.2.1.

2.3. Iterative Optimization

We evaluate our constrained decoding approach in a similar iterative optimization scheme as proposed by Nicolao *et al.* [16]. The first iteration is done on a phonetically labeled dataset to create an initial acoustic model. Afterwards, the orthographically transcribed data is labeled with automatically generated phonetic transcriptions, in the following referred to as pseudo-phonetic labels. Besides our decoding scheme described in Sec. 2, we use the external and the domain-specific extended pronunciation dictionary to create the pseudo-phonetic labels for comparison. For our decoding scheme, as it involves an acoustic model, the process can be repeated until no progress is made in terms of phone error rate (PER). We will demonstrate the performance for a small, orthographically transcribed dataset in Sec. 4.2.

3. Experimental Settings

We investigate whether our decoding method is suitable for improving inaccurate phonetic transcriptions, generated using an orthographic transcription and a pronunciation dictionary, by applying it to the kidsTALC corpus. Sec. 3.1 gives a short introduction to this corpus. We then describe details of our implementation in Sec. 3.2.

3.1. kidsTALC Corpus

We utilize the kidsTALC corpus[1] of children's speech. In this dataset we focus on monolingual German speaking children, which are typically developing and aged from 3½–11 years. In total 47 children are recorded, totaling ca. 12.6 h of unstructured, free speech. In general the elicitation context varies between story telling, picture description and conversational discourse. All audios are manually transcribed, and both the orthographic and phonetic transcriptions are checked multiple times for consistency. The dataset contains markings for unintelligible utterances, and overlapping parts. We do not consider utterances with either during the training. For more details, see [6].

To test the iterative training we use recordings of additional children, that only have been transcribed orthographically. The eligibility criteria and elicitation context is identical. For the additional data we recorded 16 children, which results in ca. 3.4 h speech. We refer to the combined corpus as kidsTALC+.

3.2. Implementation Details

As an acoustic model, we use a deep neural network (DNN) based model to compute a dense probability distribution over all phones within each timestep. Our features are Mel spectrograms computed from 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. The probability distribution over our phone set is computed by a dense layer and the softmax function. The implementation of our model is based on a recipe from SpeechBrain [22] found at *recipes/TIMIT/ASR/CTC*. To better meet the requirements of our dataset, we adjusted the learning rate ($l_r = 0.0003$), the optimizer (Adam [23]), and the learning rate scheduler (OneCycleLR [24]). To further stabilize the training we removed all augmentations and only applied frequency masking [25]. All further hyperparameters, e. g., the model structure, stay unchanged. However, as we are using a WFSA for decoding, the training loop and inference needed minor adjustments as well.

The WFSA decoding is realized with the k2 library[2]. The CTC-Topology $\mathcal{C}$ is part of this library. We use the output from our model described above directly to generate the emission graph $\mathcal{H}$. Required computations, as the composition, pruning, or the Viterbi path, are also implemented in k2. The external pronunciation dictionary is generate with a G2P model [26] using data from the BAS repository [27].

We further split the train set of the kidsTALC corpus by speakers into a train (from now on just referred to as train set) and a development split. We use the development split for tuning of the models hyperparameters and the parameters for the constrained decoding. It is not used during computation of the pronunciation statistics, as described in Sec. 2.2.2, or the extended pronunciation dictionary. Parameters to be tuned are the number of pronunciation variants to use from the train set and penalties for substitutions, deletions and insertions, introduced by the data-driven modelling as well as by modelling of PEPs.

4. Results and Discussion

In the following sections we first evaluate our constrained decoding and then give a short outlook over it's effect on using it during iterative training with orthographically transcribed data. All results are averaged over two random seeds as well as two different train/development splits. The hyperparameters are tuned once and kept consistent for all runs.

4.1. Decoding

In this section we report the effect, which the different strategies for pronunciation variation modelling presented in Sec. 2 have on the generation of a pseudo-phonetic transcription for the test set of the kidsTALC corpus. The results are shown in Tab. 1.

[1] https://www.tnt.uni-hannover.de/project/talc
[2] https://github.com/k2-fsa/k2

Table 1: *Phone error rate on the kidsTALC test set. We compare the external and the extended dictionary to different settings of our decoding scheme described in Sec. 2. In the last column, the gain relative to just taking the most likely pronunciation from our extended pronunciation dictionary is given in percent. All results are averaged over two random seeds as well as two different train/development splits.*

		PER	Relative
dictionary	external	13.91	+ 48.4 %
	extended	9.37	0 %
constrained decoding	pron variants	8.26	-11.9 %
	+ sub/del/ins	8.06	-14.0 %
	+ PEP	8.31	-11.4 %
	sub/del/ins	8.53	-8.9 %
	+ PEP	8.58	-8.4 %
	PEP	9.42	+0.6 %

4.1.1. Domain-Specific Pronunciation Dictionary

We first compare the external to the extended dictionary, which is based on the train set as described in Sec. 2.2.1. Not considering the domain-specific information increases the PER by 48.4 %. For children's speech this is expected as common pronunciations often deviate from the standard pronunciation of adult speech. Therefore, we take the results of the domain-specific pronunciation dictionary as a baseline and compare our constrained decoding approach to it in the following sections.

4.1.2. Multiple Pronunciation Variants

In this section we discuss the effect of inserting the three most common pronunciation variants of the extended pronunciation dictionary into the sequence graph S, as described in Sec. 2.2.1. The decoding will now return a less probable variant, if the model is highly certain that this specific pronunciation was present in the audio. We improve the PER relatively by 11.9 % compared to just taking the most common pronunciation using this decoding method.

4.1.3. Data-Driven Substitutions, Deletions and Insertions

Incorporating deviations from the standard pronunciations, as described in Sec. 2.2.2, reduces PER by relative 8.9 %. However, the improvements are smaller than using existing pronunciation variants. This can be explained by the fact that the elicitation context is similar for all children and therefore a large overlap in vocabulary exists between train and test set. When the data contains more diverse elicitation settings, we expect modelling on mono- to tri-phone level, instead of on a word level, will be more effective. Furthermore, the gains from using multiple variants and from modelling substitutions, deletions and insertions do not add up entirely but increase the improvement to relative 14 %. This is expected since both strategies are based on the pronunciation statistics from the training data.

4.1.4. Phonological Error Patterns

Modelling PEPs during decoding does not result in improved PER. One cause of this result possibly lies in the fact that the corpus used for this work only includes typically developed children over 3½ years old and thus PEP are already scarce for German-speaking children of this age [28]. However, modelling PEPs should be reassessed when dealing with speech of children with speech sound disorders or of younger children. The potential of allowing PEP during decoding can be seen in our results, when only evaluating on lisp sounds (/ð/ and /θ/), which do not exist in correct German speech and which are the most common mispronunciations in the kidsTALC corpus. About one third of these are correctly identified by the constrained decoding, when PEPs are modelled.

4.2. Iterative Training

In this section we discuss the effect the improved pseudo-phonetic transcription has on the training of a new ASR model. The results are shown in Tab. 2. Adding the additional data by just translating the orthographic transcription using the standard pronunciation of an external dictionary already improves the models performance in terms of PER by relative 4.6 %. Using domain-specific pronunciations further improves this by relative 1.5 % which is increased to relative 2.9 % when the constrained decoding presented in this work is applied.

For the experiments presented here, only about 30 % additional data with only an orthographic transcription was available. However, this is already enough to demonstrate the benefits of improving the pseudo-phonetic transcription for training. We expect pronunciation modelling to get more important when a larger part of the training data is not phonetically transcribed.

Table 2: *Phone error rate for unconstrained greedy CTC decoding on the kidsTALC test set. We report results for models trained with and without extra data, and compare the different approaches how the pseudo-phonetic transcription is generated for this data. All values are averaged over two random seeds and two different train/development splits used during training.*

data	transcript for extra data	PER	Relative
kidsTALC	–	25.41	+4.4 %
kidsTALC+	external dict	24.24	0 %
	extended dict	23.87	-1.5 %
	constrained decoding	23.53	-2.9 %

5. Conclusions

In this paper we presented a CTC-Decoding scheme based on WFSA, which incorporated prior knowledge from an orthographic transcription and an acoustic model, to improve the phonetic transcription of the audio. Even when the baseline pronunciation dictionary already includes domain information, i. e., pronunciations from the train set, our decoding scheme further improves the phonetic transcript by relative 14 % PER. This is achieved by using the most relevant pronunciations occurring in the train set, incorporate common substitutions, deletions and insertions by a data-driven approach, and re-scoring the weights of the WFSA using a CTC based acoustic model. Modelling PEPs, to account for pronunciations that are not present in the dataset, but common among children, has minor effects on the phonetic transcriptions. However, the used dataset only contains typically developed children in a similar setting. We expect this to be more important when more diverse settings and children with speech sound disorders are included in the data.

6. References

[1] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[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] 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.

[4] 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.

[5] 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.

[6] L. Rumberg, C. Gebauer, H. Ehlert, M. Wallbaum, L. Bornholt, J. Ostermann, and U. Lüdtke, "kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech," in *Proceedings INTERSPEECH 2022 – 23nd Annual Conference of the International Speech Communication Association*. ISCA, 2022.

[7] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-Light: A Benchmark for ASR with Limited or No Supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, 2020, pp. 7669–7673.

[8] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative Pseudo-Labeling for Speech Recognition," in *Proceedings INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020.

[9] A. Baevski, H. 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 33 (NeurIPS 2020)*, 2020.

[10] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition," *arXiv:2010.10504 [cs, eess]*, 2020.

[11] J. Olcoz, O. Saz, and T. Hain, "Error Correction in Lightly Supervised Alignment of Broadcast Subtitles," in *Proceedings INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*. ISCA, 2016, pp. 2110–2114.

[12] J. Fainberg, O. Klejch, S. Renals, and P. Bell, "Lattice-based lightly-supervised acoustic model training," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*. ISCA, 2019.

[13] E. Yılmaz, J. Pelemans, and H. Van hamme, "Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model," in *Proceedings INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*. ISCA, 2014, pp. 969–972.

[14] L. Ward, A. Stefani, D. Smith, A. Duenser, J. Freyne, B. Dodd, and A. Morgan, "Automated Screening of Speech Development Issues in Children by Identifying Phonological Error Patterns," in *Proceedings INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*. ISCA, 2016, pp. 2661–2665.

[15] E. Fringi, J. F. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proceedings INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*. ISCA, 2015, pp. 1621–1624.

[16] M. Nicolao, M. Sanders, and T. Hain, "Improved Acoustic Modelling for Automatic Literacy Assessment of Children," in *Proceedings INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*. ISCA, 2018, pp. 1666–1670.

[17] W. Chu, Y. Liu, and J. Zhou, "Recognize Mispronunciations to Improve Non-Native Acoustic Modeling Through a Phone Decoder Built from One Edit Distance Finite State Automaton." in *Proceedings INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020, pp. 3062–3066.

[18] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[19] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 167–174.

[20] M. Mohri, F. Pereira, and M. Riley, "Speech recognition with weighted finite-state transducers," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 559–584.

[21] A. Fox-Boyer, *Kindliche Aussprachestörungen: phonologischer Erwerb, Differenzialdiagnostik, Therapie*, 7th ed. Schulz-Kirchner Verlag, 2016.

[22] 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.

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

[24] 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.

[25] 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.

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

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

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