INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploiting Diversity of Automatic Transcripts from Distinct Speech Recognition Techniques for Children's Speech

Christopher Gebauer[†,1], *Lars Rumberg*[†,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

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

Abstract

The recent advances in automatic speech recognition (ASR) technologies using end-to-end machine learning do not transfer well to children's speech. One cause is the high pronunciation variability and frequent violations of grammatical or lexical rules, which impedes the successful usage of language models or powerful context-representations. Applying these methods affects the nature of the resulting transcript rather than improving the overall recognition performance. In this work we analyze the diversity of the transcripts from distinct ASR-systems for children's speech and exploit it by applying a common combination scheme. We consider systems with various degree of context: Greedily decoded and lexicon-constrained connectionist temporal classification-models, attention-based encoder decoders, and Wav2Vec 2.0, a powerful context-representation. By exploiting their diversity we achieve a relative improvement of 17.8 % on phone recognition compared to the best single system.

Index Terms: speech recognition, children's speech, model combination

1. Introduction

The advances over the past decade in automatic speech recognition (ASR) technologies usually attributes to huge end-to-end (E2E) models providing context-rich representations [1]. Concurrently, the implicit [2] or explicit [3] usage of a language model (LM) embeds additional language-specific knowledge into the resulting transcripts. Even tough this leads to an incredible recognition performance, the transfer to low-resource domains like children's speech remains a challenging task [4]. One cause for the challenges in children's speech is the high pronunciation variability and frequent violations of grammatical or lexical rules [5], which the transcript has to capture if child speech assessment is a desired downstream task [6]. Embedding highly specialized LMs can help in some cases [4], but does not transfer well between different domains of children's speech [7]. Explicitly modeling the most common error patterns is also promising [8], but requires the knowledge of a canonical pronunciation [9, 10] and, therefore, does not scale well to spontaneous speech.

This leaves us with two extrema: Either specializing a LM as far as possible towards children's speech and utilizing context-rich representations or omitting inter-frame dependencies to capture more accurately local deviations. Specializing a LM gives reasonable constraints on the decoding graph when the intelligibility of the audio segment is limited. On the other side these models capture less pronunciation variability, e. g., sequence-to-sequence (S2S), a system that intrinsically learns a LM, has problems with rare or unseen words [2]. One cause we noticed is that S2S tends to generate whole words, while avoiding word fragments or neologism. Relying on context, i. e., incorporating information from the surrounding audio to infer the local segment [1], has often a similar effect. As we will show, the combination of both increases the general robustness of the recognition system but fails to capture local variability of the children's speech. For the other extreme, capturing as much local pronunciation variability as possible, connectionist temporal classification (CTC) [11] outperforms S2S on children's speech as shown by Shivakumar *et al.* [7]. Here the omission of the inter-frame dependencies of CTC is valuable, as it captures more reliably deviations from an expected pronunciation [12, 7].

Therefore, in this work we will investigate the diversity of modern E2E-systems for children's speech and exploit their difference to improve upon phone error rate (PER). This allows us to utilize the best from both worlds: Context-dependent systems, like Wav2Vec 2.0 [1], recovering well on more difficult audio segments and context-independent systems, like greedily decoded CTC-models, that capture a higher degree of local variability. For the merging process we rely on ROVER [13], a method that is commonly applied for such tasks. Knill *et al.* [14] also combine transcripts for children's speech using ROVER. However, the authors neither considered E2E-systems nor investigated the nature of the transcripts. Former is of special interest, as Shivakumar *et al.* [7] demonstrated that E2E models outperform HMM-DNN systems on children's speech in terms of character and word recognition. Kurata *et al.* [15] proposed a powerful merging scheme for CTC-based systems. We will show that not the merging scheme, but the diversity of the utilized transcripts is more important. Our contribution is two-fold: We analyze different modern ASR-systems on children's speech and summarize their strengths as well as weaknesses. Furthermore, we quantify the effect of combining the most diverse systems in terms of PER and analyze the effect on child speech assessment, a possible downstream task.

2. Speech Recognition Techniques

In this section we describe the investigated ASR-techniques. In the following we refer to the features of a small audio window as *frame*. The *token* represents a character or phone, depending on the desired output sequence. We define a *system* as a combination of training techniques and feature extractors, while a *model* defines a specific trained instance of a system.

† contributed equally

10.21437/Interspeech.2023-926

2.1. Context-Independent Techniques

In this work we consider models trained with connectionist temporal classification (CTC) [11] as context-independent due to the omission of the inter-frame correlation. This assumption does not strictly hold, because convolutional neural networks (CNNs) provide information from the surrounding frames. Nevertheless, we show in Sec. 5 that CTC has the best capabilities to capture local deviations from an expected pronunciation compared to, e. g., S2S.

CTC solves the frame-token alignment problem efficiently using dynamic programming. First, all consecutive, identically decoded time-frames are collapsed to one token. An additional <*blank*>-token allows the repetition of tokens, which is removed in a second step. The simplest and also most independent way of decoding such systems is to greedily select the Viterbi path, which we will refer to as CTC-greedy. A more elaborated way of decoding is described in Sec. 2.2.1

2.2. Context-Dependent Techniques

In this section we will describe two different systems that constrain the resulting output, either explicitly [3] or implicitly [2]. Furthermore, we utilize Wav2Vec 2.0 (W2V) [1] in combination with all presented training/decoding techniques. It is known to provide a powerful context-representation.

2.2.1. Constrained CTC Decoding

Constraining the CTC-emissions has a lucrative advantage: No additional system needs to be trained. One approach for constrained CTC decoding is to utilize weighted finite state transducer and efficiently limit the dense emission graph $\mathcal{E}$ from the CTC-model by composing it with a lexicon and LM [3]:

$$\mathcal{D} = \mathcal{E} \circ (\mathcal{C} \circ \min(\det(\mathcal{L} \circ \mathcal{G}))),$$

where $\mathcal{D}$ is the resulting decoding graph, $\mathcal{L}$ a lexicon, and $\mathcal{G}$ a LM. $\mathcal{C}$ is a CTC-topology allowing for repetitions of characters and insertions of the blank token. The operator $\circ$ represents the composition of the graph, min the minimization and det the determinization, respectively. Both, the lexicon and the LM, are commonly trained on large amounts of in domain text-data, where a corresponding audio is not necessarily required. In our work we will omit the LM and only constrain our decoding towards a lexicon that is based on in-domain samples from a children's speech corpus.

2.2.2. Sequence-to-Sequence

Sequence-to-sequence systems intrinsically learn a LM [2], which makes them especially powerful if large amounts of data are available. In this work we utilized an attention-based encoder decoder (AED) system [2]. A bidirectional RNN-encoder computes for each frame a feature vector, which are combined by building a weighted sum. In our work the weights are computed based on a location-based attention. Including the previously decoded token into the model input is the major cause that AEDs implicitly learn a LM and lexicon.

Watanabe *et al.* [16] introduced a hybrid CTC/S2S training and decoding scheme. Not including the CTC-loss into the training process lead to divergence in our case and, therefore, we do not further investigate the impact of plain S2S training or its decoding. Additionally, we noticed a higher diversity towards the other systems when using the algorithm described by Watanabe *et al.* [16] compared to greedily decoding the S2S-model, which is desirable in our work.

2.2.3. Wav2Vec 2.0

Wav2Vec 2.0, introduced by Baevski *et al.* [1], is a context-based feature representation of raw audio. The network consists of three main components: A CNN-based feature extractor, a transformer-based context network, and a quantized codebook. The entire system is trained E2E in an unsupervised fashion, which allows to utilize large amounts of unlabeled data. The general idea of the training procedure is that the context network has to recover randomly masked information from neighboring features. The resulting model provides a strong context-representation for speech, which is commonly finetuned by training a small dense head in supervised fashion using the methods introduced in Sec. 2.1 and Sec. 2.2.2.

3. Model Combination

In this section we discuss, how an ensemble of N transcripts is merged. The ensemble may consists of transcripts from models of the same ASR-system or from different systems.

3.1. ROVER

As mentioned, we rely on ROVER [13], a simple but effective system for transcript combination. First, the N transcripts are globally aligned, where the alignment method is independent of ROVER. In the original work two different versions are proposed to select the final token per index in the global alignment: Weighted selection based on the uncertainty of the underlying ASR-system and frequency based selection. While a token-level uncertainty is naturally given for S2S-based models, the usage of CTC-models makes an approximation necessary. In the next section we will introduce different weighting schemes for ROVER and evaluate their general necessity in Sec. 6.

3.2. Weight Selection

CTC-based systems do not provide an uncertainty measure on token-level intuitively. While there exists a method based on Monte Carlo Dropout (MCD) [17], we did not find any improvements in terms of recognition performance using them for weighting (see Sec. 6). This is reasonable as the same ensemble is applied to ROVER.

Instead, we directly target phonetic inventories, a possible downstream task commonly applied during child speech assessment [6], by manually selection suitable weights. Phonetic inventories require high accuracy especially on rare phones. We show in Sec. 6 that manually increasing the weighting for one specific system, which achieves a greater performance on rare phones, already leads to higher accuracy for these phones without harming the accuracy of the other phones.

4. Experimental Setting

We train and evaluate all models on the kidsTALC corpus [18], which consists of typically developing, monolingual, and German speakers aging from 3 ½–11 years. The training data is based on ∼5 h of spontaneous children's speech, equally distributed over gender and age. For further information on the characteristics of the kidsTALC corpus we refer to the corresponding publication [18]. As suggested by the authors, we extend the kidsTALC corpus by Mozilla Common Voice (MCV) [1]. We vary the presence of MCV by adjusting the fraction ρ_{MCV}

[1] `https://commonvoice.mozilla.org/en/datasets`, Version 11.0, German

in each batch between $\rho_{MCV} \in \{0.25, 0.5\}$, treating it as a hyperparameter. From MCV the model sees a maximum of $\sim$465 h of speech during training, due to the huge difference in corpus size each segment is only seen once.

We implement and train all models based on the TIMIT recipe[2] for ASR from SpeechBrain [19] and deploy for all decoding schemes the default settings. In our work the model output represents a probability distribution across the restricted IPA phone set proposed in the kidsTALC corpus [18]. For stability reasons, we adjust the recipe and deploy the Adam optimizer [20] and the OneCycleLR scheduler [21]. We adjust the learning rate $l_r \in \{1e-3, 3e-3\}$ and dropout rate $d_r \in \{0.15, 0.25\}$, respectively. By also varying the random seed we obtained multiple models and selected the best five different models per system using the validation split. Additionally, we average all results on the test set using two different train-validation splits. For the constrained CTC decoding we adopt k2[3]. We reimplement ROVER using the lingpy package [22] for the global alignment.

Regarding W2V, we use the MCV recipe[4] for ASR from SpeechBrain [19] with the same learning rate and scheduler as above. We train a small dense network on kidsTALC+MCV using the output of a pretrained W2V-model from Hugging Face[5] as input features. The feature extractor as well as the context network are frozen. This setup replaces the Mel feature banks combined with a CRDNN as applied in the basic recipes above. We applied the same training approaches from Sec. 2.1 and Sec. 2.2.2.

5. System's Diversity

In this section we will investigate the diversity of all trained systems. We will start by comparing the systems to the manual transcript, both on the phone error rate (PER) and the word error rate (WER). The WER is computed on the phonetic and not on the orthographic transcript. Even tough we are exclusively interested in the improvement of the PER, we notice the WER to be helpful when analyzing the diversity on a coarser level. For the combination of W2V and S2S, we notice some models to completely diverge, i. e., reaching extremely high error rates. If the PER is greater than 100.0 on the validation set, we do not consider the models in this section. However, they are included in Sec. 6 as they still contribute positively to the final results. Next we compute the difference between all systems, again on PER and WER. Closing, we analyze if some systems are more performant in recognizing certain phones than others.

In Tab. 1 we compute the PER and WER with respect to the manual transcript. The systems based on W2V always outperform their direct counterparts. The same holds for ROVER, when joining multiple models of one system. Generally, CTC-based systems, when greedily decoded, outperform all other systems with respect to PER. On the other hand, S2S-based systems outperform the other systems with respect to WER. This shows, that omitting the inter-frame dependencies helps to be locally more accurate, but generally fails more often to capture the entire word. Overall all systems perform similarly well, which is interesting as we noted high disagreement between the resulting transcripts, which we will discuss next.

[2] https://github.com/speechbrain/speechbrain/tree/develop/recipes/TIMIT/ASR, commit 1bc762c

[3] https://github.com/k2-fsa/k2, version 1.23.2

[4] .../recipes/CommonVoice/ASR, commit 002779c

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

Table 1: *PER and WER with respect to the manual transcript. Avg refers to the mean across all models after computing the error rate, while [13] refers to merging the transcripts using ROVER first. For any speech recognition system ROVER and W2V works best. CTC-greedy is better on PER, while S2S performs better on WER. Con. represents CTC-Constrained.*

		CRDNN			W2V+DNN		
		Greedy	Con.	S2S	Greedy	Con.	S2S
PER	Avg	23.7	24.4	42.7	21.0	21.5	25.3
	[13]	**22.1**	22.9	31.4	**19.7**	20.7	21.5
WER	Avg	61.5	61.1	59.7	55.5	55.6	50.7
	[13]	58.7	59.8	**54.5**	52.9	54.6	**47.8**

Table 2: *Difference between all utilized systems using the PER. We averaged across all seed combinations for two given systems. Gr. represents CTC-greedy and Con. CTC-Constrained.*

		CRDNN			W2V+DNN		
		Greedy	Con.	S2S	Greedy	Con.	S2S
CRDNN	Gr.	19.0	-	-	-	-	-
	Con.	26.6	18.9	-	-	-	-
	S2S	42.0	45.4	39.7	-	-	-
W2V+D	Gr.	34.1	36.6	50.3	16.5	-	-
	Con.	34.9	32.2	50.4	23.8	15.5	-
	S2S	36.1	38.0	47.4	24.3	27.3	15.8

In Tab. 2 we computed the PER between two given systems, not considering the manual transcript. As mentioned, the disagreement between the systems is higher than the error rate towards the manual transcript, i. e., merging these systems is promising. Tab. 2 shows that the similarity between models of a given system (diagonal values) is always higher than to any other system. This means that the inclusion of multiple distinct systems increases the diversity in the resulting ensemble and, therefore, the potential for improvement. Going one step further, we find more similarity when grouping the systems by the feature encoder, i. e., if the considered systems are based on a CRDNN (top left section) or W2V+DNN (bottom right section). Intuitively, that makes sense, because W2V introduces a high amount of context into the ASR-system. Therefore, the highest dissagreement is between CRDNN and W2V-based systems (bottom left section). It needs to be noted that the combination of CRDNN+S2S greatly varies from any other combination, even leading to high discrepancy between its models.

This behavior is not visible, when looking at the same table for WER, as shown in Tab. 3. The combination of CRDNN+S2S has a intra- and inter-system consistency of similar magnitude compared to the other systems when evaluating on the WER. Increasing the amount of context and LM-constraints improves the WER on intra-system agreement, bringing it closer to the magnitude of the PER in Tab. 2. This means models from a system relying on W2V and LM-constraints mainly output either entirely identical or completely different words. This shows again that these system are not able to capture local variability well.

In our work we define the performance of a system on a certain phone by the difference between true positives and false negatives, i. e., correctly labeled phones minus its insertions. We refer to this metric as phone goodness. In Tab. 4 we list

Table 3: *Difference between all utilized systems using the WER. We averaged across all seed combinations for two given systems. Gr. represents CTC-greedy and Con. CTC-Constrained.*

		CRDNN			W2V+DNN		
		Greedy	Con.	S2S	Greedy	Con.	S2S
CRDNN	Gr.	57.0	-	-	-	-	-
	Con.	68.8	39.8	-	-	-	-
	S2S	67.8	66.6	54.6	-	-	-
W2V+D	Gr.	77.2	77.3	77.7	50.7	-	-
	Con.	76.6	61.3	72.3	64.2	33.0	-
	S2S	73.8	70.5	67.8	58.6	57.8	35.5

Table 4: *List of phones on which a system outperforms the other systems on phone goodness by at least 5 %. Gr. represents CTC-greedy and Con. CTC-Constrained.*

CRDNN	Gr.	/t/, /ð/, /θ/, /χ/
	Con.	/ˌ/, /χ/
	S2S	
W2V+D	Gr.	/m/, /p/, /t/
	Con.	/ˌ/, /uː/, /oe/
	S2S	/ˌ/, /g/, /h/, /m/, /p/, /t/, /uː/, /ŋ/, /ɐ/, /ʏ/

all phones of a system for which the phone goodness outperforms the next best system by at least 5 %. We allow up to three equally well performing systems for one phone if they share a gap to the next best performing system. CTC-greedy without W2V generally performs better on fricatives, but especially well on /ð/, /θ/, /χ/. Systems based on W2V are usually better on nasals and plosives, but also on certain vowels.

6. Application

Before we present the results of the model combination, we analyze the improvement in terms of phone recognition by applying uncertainty for ROVER and compare ROVER with the merging scheme of Kurata *et al.* [15]. Since the uncertainty for CTC-based systems is most difficult, we evaluate the uncertainty's impact on PER only for the system-combination of CTC-greedy and CTC-constrained. Applying the unweighted ROVER leads to a PER of 20.8 %, while weighting the ROVER using MCD [17] results in 21.6 %. Therefore, we do not consider any uncertainty estimation for ROVER. To compare ROVER with Kurata we merge different models of CTC-greedy. For ROVER this leads to 22.1 %, as seen in Tab. 1, and Kurata achieves 22.4 %. While Kurata improves upon the unmerged models, it is slightly worse than ROVER.

In Tab. 5 we combine all systems with the same feature encoder. We show that the relative improvement, towards the single best system joined with ROVER, is independent of the feature encoder. The largest gain is achieved if all three systems are combined, leading to 10.0 % and 10.7 % relative improvement for CRDNN and W2V, respectively. When only considering two systems, CTC-greedy and S2S perform better compared to CTC-greedy and CTC-constrained. This behavior is expected based on the diversity analysis in Sec. 5. However, latter has the benefit that no additional model needs to be trained.

In Tab. 6 we also combine the different feature encoders, as Sec. 5 showed a higher diversity than sticking to one representation. Merging all systems does perform already much better.

Table 5: *PER with respect to the manual transcript. We combine different ASR-systems with the same feature encoder using ROVER [13]. Merging all systems works best. Con. represents CTC-Constrained.*

Greedy	Con.	S2S	CRDNN	W2V+DNN
✓	X	X	22.1	19.7
✓	✓	X	20.8	18.6
✓	X	✓	20.4	17.9
X	✓	✓	20.7	18.5
✓	✓	✓	**19.9**	**17.6**

Table 6: *PER with respect to the manual transcript. We combine different ASR-systems using ROVER [13], also combining different feature encoder. Due to the higher diversity this reasonably leads to the overall best results.*

CRDNN			W2V+DNN			PER
Greedy	Con.	S2S	Greedy	Con.	S2S	
✓	✓	✓	✓	✓	✓	16.2
✓	X	✓	✓	X	✓	**16.2**
✓	X	X	X	X	✓	17.3

However, an ablation study showed that CTC-constrained is not necessary. Overall this leads to a PER of 16.2 %. Compared to the best single-system combination including ROVER, which is W2V+CTC-greedy with a PER of 19.7 % (see Tab. 1), we achieve a relative improvement of 17.8 %.

A phonetic inventory is a possible downstream task for children speech assessment. This makes especially the performance on the rare phones important. While the weighting according to an uncertainty was not promising, we manually set a weight for certain systems. In our work we exemplary target to improve on the fricatives and, therefore, increase the weight for the fricative outputs of all CTC-greedy models. For example this trivial approach already corrects 10.8 % with respect to the best combination in Tab. 6 and 54.1 % with respect to the best single model in Tab. 1 of the incorrect labels of the phone /ç/. We will investigate more elaborated methods of weighting and the direct impact on the downstream task in future work.

7. Conclusion

In this paper, we investigate the diversity as well as the strengths and weaknesses of modern E2E systems on children's speech. We notice high dissimilarities in the resulting transcripts on token and word level between the investigated systems, while maintaining a similar overall performance with respect to a manual transcript. Applying powerful context-representations and LM-constraints has a high impact on the nature of the transcript, which results in an improved WER while loosing the local variability. We exploit this diversity by applying ROVER, a common system for transcript merging. Fully exploiting the divers set of systems leads to a relative improvement of 17.8 % on the PER.

8. References

[1] 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 (NeurIPS)*, 2020.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[3] 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.

[4] J. Thienpondt and K. Demuynck, "Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping," in *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*. ISCA, 2022, pp. 2213–2217.

[5] 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.

[6] T. M. Byun and Y. Rose, "Analyzing Clinical Phonological Data Using Phon," *Seminars in Speech and Language*, vol. 37, no. 2, pp. 85–105, 2016.

[7] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[8] 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.

[9] L. Rumberg, C. Gebauer, H. Ehlert, U. Lüdtke, and J. Ostermann, "Improving Phonetic Transcriptions of Children's Speech by Pronunciation Modelling with Constrained CTC-Decoding," in *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*. ISCA, 2022, pp. 1357–1361.

[10] 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.

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

[12] X. Yue, G. Lee, E. Yılmaz, F. Deng, and H. Li, "End-to-End Code-Switching ASR for Low-Resourced Language Pairs," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 972–979.

[13] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 1997, pp. 347–354.

[14] K. M. Knill, L. Wang, Y. Wang, X. Wu, and M. J. Gales, "Non-Native Children's Automatic Speech Recognition: The INTERSPEECH 2020 Shared Task ALTA Systems," in *Proceedings INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020, pp. 255–259.

[15] G. Kurata and K. Audhkhasi, "Guiding CTC Posterior Spike Timings for Improved Posterior Fusion and Knowledge Distillation," in *Proceedings INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association*. ISCA, 2019, pp. 1616–1620.

[16] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[17] A. Vyas, P. Dighe, S. Tong, and H. Bourlard, "Analyzing Uncertainties in Speech Recognition Using Dropout," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6730–6734.

[18] 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 – 23rd Annual Conference of the International Speech Communication Association*. ISCA, 2022, pp. 5160–5164.

[19] 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, commit 5df3885c. [Online]. Available: https://github.com/speechbrain/speechbrain

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

[21] 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.

[22] J.-M. List and R. Forkel, "LingPy. A Python library for historical linguistics," Max Planck Institute for Evolutionary Anthropology, 2021.