INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Uncertainty Estimation for Connectionist Temporal Classification Based Automatic Speech Recognition

Lars Rumberg[†,1], *Christopher Gebauer*[†,1], *Hanna Ehlert*[2], *Maren Wallbaum*[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

Predictive uncertainty estimation of deep neural networks is important when their outputs are used for high stakes decision making. We investigate token-level uncertainty of connectionist temporal classification (CTC) based automatic speech recognition models. We propose an approach, which considers that not all changes at frame-level lead to a change at token-level after CTC decoding. The approach shows promising performance for prediction of recognition errors on TIMIT, Mozilla Common Voice (MCV) and kidsTALC, a corpus of children's speech, using two different model architectures, while introducing only negligible computational overhead. Our approach identifies over 80 % of a wav2vec2.0 model's errors on MCV by selecting 10 % of the tokens. We further show, that the predictive uncertainty estimate relates to the uncertainty of a human annotator, by re-annotating 500 utterances of kidsTALC.

Index Terms: Uncertainty, Automatic Speech Recognition, Children's speech

1. Introduction

As for many other applications, deep neural networks (DNNs) have become the prevailing approach for automatic speech recognition (ASR). When they are used in any decision making, errors can have serious implications. For example, decisions of ASR systems used for therapeutic assessment of children's speech and language development can have large consequences for the children's further development [1, 2]. Predictive uncertainty estimation of DNNs is therefore a highly relevant topic.

In traditional Hidden Markov Model based ASR systems, an uncertainty can be estimated using the scores contained in word lattices [3, 4]. For end-to-end DNN based ASR systems this is not as straight forward. Some prior approaches train external models to predict the uncertainty [5, 6]. Uncertainty estimation of end-to-end ASR without external models, which is the focus of this work, is investigated in [7, 8, 9].

Malinin *et al.* [7] derive multiple information-theoretic measures of uncertainty for autoregressive models using deep ensembles [10], and evaluate them on machine translation and ASR. Oneaţă *et al.* [8] also work with autoregressive models. They compare the use of the output's entropy to using the log-probability of the most probable token for token-level uncertainty, as well as multiple aggregation methods to compute word-level uncertainty. They further show that temperature scaling and the usage of ensembles, both Monte Carlo Dropout (MCD) ensembles [11] or an ensemble of independently trained networks [10], can improve the estimation.

While autoregressive models have shown great performance for adult speech recognition, they have been shown to be less effective for child speech. Shivakumar *et al.* [12] show in a large empirical study that, on child speech, autoregressive models are outperformed by connectionist temporal classification (CTC) [13] models. They also show that language model re-scoring is often ineffective for child speech. Each token-level decision of an autoregressive model is conditioned on all preceding tokens in the sequence. This enables them to learn an implicit language model, explaining in part their good performance on adult speech [14]. Given the high linguistic variability of child speech, the assumption of conditional independence between time-frames of CTC models seems to be preferable.

While both the approaches of [7] and [8] can in theory be used with non-autoregressive models based on CTC, this would neglect that a part of a CTC-based model's uncertainty is about the alignment of the token sequence to the input and not about the token sequence itself. We show, that by considering that some changes in the frame-level output do not lead to a different token-sequence after decoding, the token-level uncertainty estimates of a CTC-based model can be improved.

Vyas *et al.* [9] also estimate the uncertainty of CTC-based ASR models. They compute the disagreement between the decoded predictions of a MCD ensemble. A similar approach is also used in [15] to improve self-training for domain adaptation for ASR. By using the decoded predictions, they are robust to any uncertainty about the alignment. However this approach relies on an (MCD-) ensemble and discards all information about the models uncertainty in the output probabilities.

In this work we compare different approaches for uncertainty estimation of CTC-based ASR models. We present a method, which takes into account, whether changes on a frame-level would lead to a different decoded output. We evaluate the methods for phone recognition on TIMIT [16], and kidsTALC, a publicly available child speech corpus [17], as well as for orthographic speech recognition on Mozilla Common Voice (MCV). We show the performance over two different model architectures, a simple CNN-RNN based model, as well as wav2vec2.0 [18], both trained, respectively fine-tuned using the CTC criterion. We further investigate, using re-annotations of kidsTALC, how the model uncertainty relates to uncertainty of human annotators.

2. Method

In the following sections we will first go into more detail on what to consider when analyzing output probabilities of CTC based ASR models for uncertainty. We propose our method for token-level uncertainty estimation and describe different approaches of how to incorporate ensembles of models.

[†] contributed equally

10.21437/Interspeech.2023-907

2.1. CTC-aware token-level uncertainty

Connectionist temporal classification (CTC) [13] defines the probability of a token sequence as the sum of the probabilities of all possible alignments between that token sequence and the audio. For a sequence level uncertainty estimation which takes all these alignments into account, one can use the CTC-loss of the decoded prediction. Frame-level uncertainty measures such as the probability of the most probable token, or the entropy over the token distribution only estimate the uncertainty of the most probable alignment.

During CTC decoding, first a frame-wise greedy decision is done by taking the argmax of the frame's probability distribution $p(y_t)$. Then, identical consecutive tokens are collapsed. Given the decisions of the frame $\hat{y}_t$ and its neighboring frames $\hat{y}_{t-1}, \hat{y}_{t+1}$, the function

$$f(y_t^*) = \begin{cases} & y_t^* = \hat{y}_t \\ 0 & \vee\, (\hat{y}_{t-1} \neq \hat{y}_{t+1} \wedge \hat{y}_t \in \{\hat{y}_{t-1}; \hat{y}_{t+1}\} \\ & \wedge\, y_t^* \in \{\hat{y}_{t-1}; \hat{y}_{t+1}; \epsilon\}) \\ 1 & \text{else} \end{cases} \tag{1}$$

describes whether changing the decision $\hat{y}_t$ to y_t^* affects the collapsed output. In words, if the token $\hat{y}_t$ is equal to exactly one neighbor, changing it to the other neighbor or the blank does not change the decoded output. If $\hat{y}_t$ is not equal to any neighbor or if both neighbors are identical, changing it always affects the output. We define the frame-level uncertainty as

$$p_{\text{change}} = \sum_{j \in \mathcal{T}} p(y_t = j) * f(j), \tag{2}$$

with $\mathcal{T}$ being the set of all tokens. In words, p_{change} is the sum of the probabilities of all tokens that would change the output sequence when one of them is decoded at that frame. For aggregation of the frame-level uncertainties to token-level we take the uncertainties of all consecutive frames with the same most probable token and compare computing the mean, minimum and maximum of those. Uncertainties of the blank token are added to both neighboring tokens.

We compare our suggested approach with computing the frame-level uncertainty, using the probability of the most probable token, as $1 - \max(p(y_t))$ and aggregating to token-level in an identical fashion. Further we compare to the method presented in [9], which we describe in the following section, as it requires an ensemble.

2.2. Ensembles

Bayesian inference can be approximated by treating an ensemble of models as sampled from an approximate posterior distribution over the parameters given the training data. The predictive posterior is computed by averaging over the ensemble's outputs. The ensemble can be generated from individually trained models [10] or a Monte Carlo Dropout (MCD) ensemble [11], where one model is trained using dropout [19]. Dropout is then being kept active during inference, which is repeated multiple times for the same sequence.

We use the latter approach and compare different model averaging approaches: First, we average the frame-level output probabilities and then compute uncertainties as described in Sec. 2.1. This approach is successfully used by [8] to improve the uncertainty estimates of autoregressive models. We argue in 4.1.2 why this might be problematic for CTC-based models. As an alternative we therefore first compute the uncertainties on token-level for each ensemble output, align the decoded outputs and then average the uncertainties on a token-level.

We further compare our approach to the one presented in [9] for error localization, adapting it to compute uncertainty estimates on a token-level instead of on word-level. For this we do inference once with dropout deactivated and use this as a reference. Then, we decode each ensemble output individually and compute the uncertainty by aligning them to the reference, counting the number of disagreements for each token and dividing this count by the amount of models in the ensemble.

3. Experimental Setting

We evaluate the proposed approach for phone recognition using TIMIT [16] and the kidsTalc corpus [17] and for orthographic speech recognition on the German part of the MCV corpus[1]. KidsTalc consists of speech of 3 ½ –11 years old German speaking children with typical speech development. In total eight hours of phonetically transcribed speech is used, from which for testing and validation each around one and a half hours is held out.

To facilitate reproducibility, we use a pretrained model of the CTC TIMIT recipe[2] of the Speechbrain toolkit [20] for the TIMIT corpus and do no training ourselves. For the kidsTalc corpus we use the same model and training setting as the baseline for the corpus suggested in [17]. It uses the same Speechbrain recipe, with small modifications, and includes the German part of the MCV corpus during training to increase the diversity.

For the MCV corpus, we again use a pretrained model of Speechbrain without doing any training ourselves. To investigate how well our method transfers to different model architectures, we use a recipe using a wav2vec2.0 [18] model[3], which was fine-tuned using CTC on the German part of the MCV corpus. The TIMIT and kidsTALC models are trained with 15 % dropout rate. The MCV recipe uses 15 % and 10 % dropout rate for the fully connected layers and the encoder, respectively. For the MCD ensemble we infer with active dropout 50 times.

To investigate how the predictive uncertainty estimate relates to the uncertainty of a human annotator, we randomly select 500 utterances of kidsTalc and redo the phonetic transcription. The annotators are trained speech language therapists. For re-annotation we use the orthographic transcription and suggest for each word multiple pronunciations to choose from. If the correct transcription is not in these suggestions, the transcriber can edit one suggestion or manually add a new one. We align the new transcript to the old to identify the tokens which have been changed. When a token was inserted, we tag both neighboring tokens as changed.

4. Results and Discussion

We compare the suitability of the uncertainty estimation approaches for model error prediction and for prediction of token changes during the re-annotation described in Sec. 3.

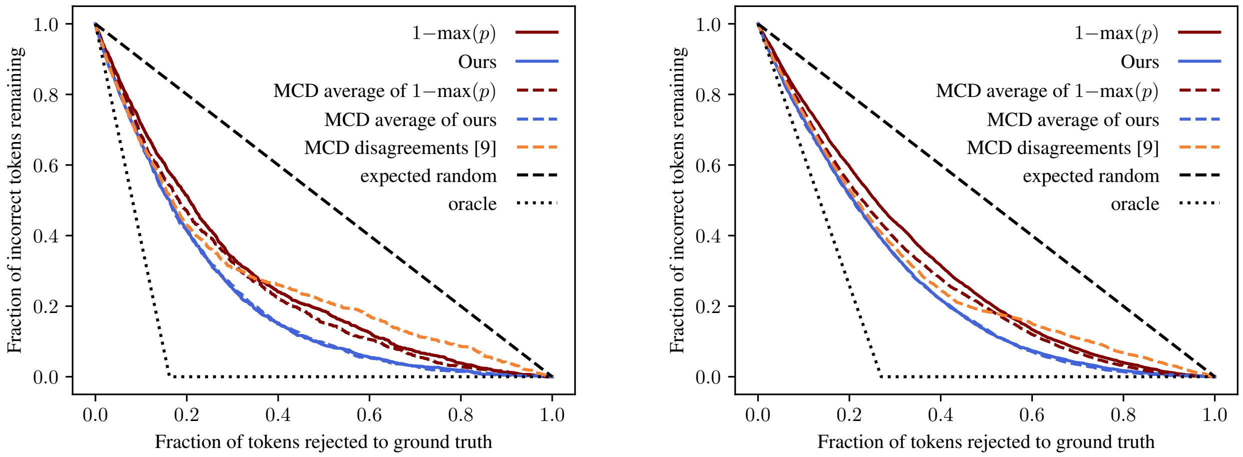
4.1. Prediction of model errors

Here, we will first introduce the used metrics and figures. We will then discuss the results of approaches using single models

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

[2] `https://github.com/speechbrain/speechbrain/tree/develop/recipes/TIMIT/ASR/CTC` Commit 0110f4e

[3] `.../recipes/CommonVoice/ASR/CTC` Commit 0110f4e



(a) TIMIT

(b) kidsTALC

Figure 1: *Prediction rejection curves for phone recognition on the test sets of TIMIT and kidsTALC. Curves close to the oracle curve indicate good prediction of errors, curves close to the expected random curve indicate random uncertainty estimates. Our method (blue) shows the best performance for error prediction.*

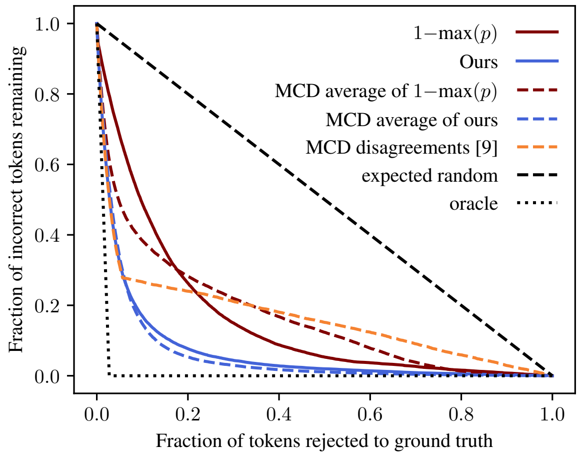


Figure 2: *Prediction rejection curves on the test set of the German MCV using a wav2vec2.0 model fine-tuned with CTC.*

in Sec. 4.1.1 before discussing the results using MCD ensembles in Sec. 4.1.2.

To evaluate suitability for model error prediction, we use prediction rejection curves. For each approach, all tokens in the test set are sorted by their estimated uncertainty. The predictions with the largest uncertainty are then rejected to the ground truth, i.e. they are replaced by their corresponding ground truth label. The amount of remaining incorrect tokens is plotted as a function of the amount of rejected tokens. Both axes are normalized by their respective total amount before rejection.

If the uncertainty estimation is random, the expected rejection curve is a straight line from the total amount of incorrect tokens in the full set to the lower right. The further a rejection curve is under this line, the better the uncertainty estimation. The rejection curves on the test sets of TIMIT and the kidsTALC corpus are shown in Fig. 1 and on MCV in Fig. 2.

We also compute the prediction rejection ratio PRR as proposed in [21]. The area between the actual rejection curve and the rejection curve of expected random uncertainty is computed and divided by the area between the curves for oracle rejection and random rejection. It is equal to 1 for perfect rejection and close to 0 if the uncertainty estimate is random. PRR on all investigated datasets is shown in Tab. 1.

Table 1: *Prediction Rejection Ratio PRR [21] on the test sets of TIMIT, kidsTALC and the German part of MCV. A value of* 1 *indicates perfect prediction of model errors using the uncertainty estimate, values close to* 0 *indicate random uncertainty estimates.*

	Frame-level uncertainty	1−max(p)		Ours		MCD disagreements [9]
	Best aggreg. to token level	min		max		
	MCD ensemble	-	✓	-	✓	✓
PRR	TIMIT	0.54	0.59	**0.69**	0.68	0.53
	TALC	0.52	0.58	0.66	**0.67**	0.57
	MCV	0.71	0.69	0.89	**0.90**	0.69

4.1.1. Single Models

For both methods working with a single model, we first evaluate how to aggregate frame-level uncertainties to token-level. Taking the maximum over frame-level uncertainties of identical consecutive tokens performed best for our method. When computing $1 - \max(p)$ on frame-level, taking the minimum gave the best results. In all figures and tables the best aggregation method for each uncertainty estimation method is used.

Comparing both methods which use the output probabilities of a single model (solid lines in Fig. 1), our method outperforms just using the probability of the most probable token, showing the importance of taking the ambiguity of alignment between audio and label sequence of CTC into account. The difference between the approaches is highest using the fine-tuned wav2vec2.0 on MCV (see Fig. 2 and Tab. 1). Here our method achieves a PRR of 0.89. Selecting the tokens with the $10\,\%$ highest uncertainty already identifies over $80\,\%$ of the model's errors. Using the naive approach of using the probability of the most probable token, this would only identify around 50 % of the model's errors, when not using an ensemble.

4.1.2. MCD Ensembles

Using a MCD ensemble, we first observe that simply averaging the output probabilities of the ensemble's models, like done by [8] for autoregressive models, does not improve the uncertainty estimation. More importantly, decoding with the averaged output probabilities also has a large effect on the decoding itself. It increases the amount of deletions by around 25 %. When different models of the ensemble align audio to label sequence even just slightly different, high but narrow probability peaks for the same token in the outputs of all models average to wide low peaks. These averaged peaks are often not high enough for the token to be decoded, leading to deletions. This again demonstrates, that uncertainty estimation approaches for autoregressive models, e. g. as presented by [8], cannot be simply applied to CTC-based models. To avoid clutter, we omit plotting the rejection curves of this ensemble averaging approach in the figures.

When instead first computing token-wise uncertainties for all models of the MCD ensemble, and then averaging those, the approach of using the probability of the most probable token benefits on TIMIT and kidsTALC. On MCV this improves the prediction of errors for high uncertainties, but worsens it for lower uncertainties, leading over all to a similar PRR (see Tab. 1). Our approach does not benefit significantly from averaging over the MCD ensemble. However it already outperforms all other evaluated approaches with just one model on all datasets, even after they are averaged over the ensemble.

Our approach also outperforms analyzing the disagreements of the decoded predictions of a MCD ensemble (orange curve in Fig. 1) as proposed in [9]. The dent in the prediction rejection curve at a rejection of around 30 % on TIMIT, 40 % on kidsTalc and 7 % on MCV of this method is a result of all models of the ensemble agreeing on the token. That means, the uncertainty is predicted to be zero for around 70 %, 60 %, respectively 93 % of the tokens of TIMIT, kidsTALC and MCV.

In general, for all approaches including a MCD ensemble it has to be regarded, that inference time linearly increases with the number of forward passes. Especially for large models like wav2vec2.0 this is an important factor for practical applications. While all results reported with a MCD ensemble needed 50 times the amount of compute for inference, the presented approach adds only a small computational overhead during decoding which is negligible compared to inference of the model.

4.2. Relation to token changes during re-annotation

We investigate the relation between the estimated uncertainty of the ASR model and uncertainty of human annotators by re-annotating 500 randomly selected utterances of the kidsTALC corpus. We plot the number of changed tokens during re-annotation as a function of the number of selected tokens, sorted by their estimated uncertainty, in Fig. 3. Equivalent to the prediction rejection ratio, we compute the ratio of the area between the actual curve and the curve of expected random uncertainty, to the area between the oracle curve and the expected random curve (Tab. 2).

All three approaches to uncertainty estimation predict token changes significantly better than random. As for prediction of model errors, our approach is the most suitable to predict token changes.

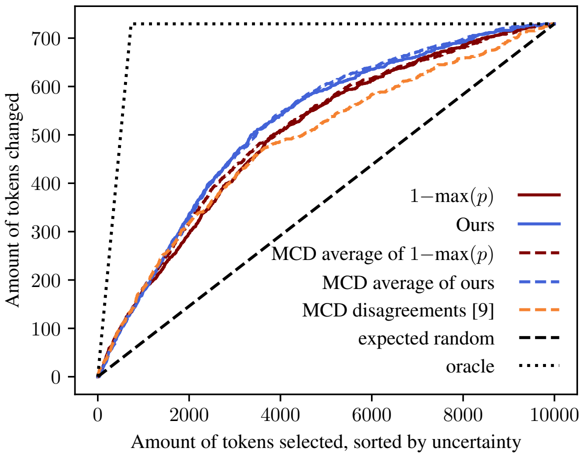


Figure 3: *Analysis of relation between uncertainty estimate and token change during re-annotation of* 500 *randomly selected utterances of kidsTalc.*

Table 2: *Area between the actual curve in Fig. 3 and the expected random curve divided by the area between the oracle curve and the expected random curve. A value of* 1 *indicates perfect prediction of token change during re-annotation using the uncertainty estimate, values close to* 0 *indicate no relation between uncertainty estimate and token change.*

Frame-level uncertainty	1−max(p)		Ours		MCD disagreements [9]
Best aggreg. to token level	min		max		
MCD ensemble	-	✓	-	✓	✓
PRR TALC	0.41	0.42	0.46	**0.47**	0.36

5. Conclusion

In this paper we presented a method for token-level uncertainty estimation of CTC-based ASR models, which considers which changes at a frame-level lead to changes at a token-level after CTC decoding. We compare it with just using the probability of the most probable token at each frame and with analyzing the disagreements of the decoded predictions of a MCD ensemble as proposed in [9]. We further average the former and our method over a MCD ensemble. On both TIMIT, kidsTALC, a phonetic German child speech corpus, and on the German part of MCV, our approach performs best for prediction of model errors, even when just using a single model. With negligible computational overhead, it allows identifying over 80 % of a wav2vec2.0 model's errors on MCV, by selecting 10 % of the tokens. Using pre-trained models without doing any training ourselves, we demonstrate the capability of the presented approach to be used with any pre-existing CTC-based model. We further show that the predicted uncertainty relates to the uncertainty of a human annotator.

Further work should investigate the suitability of the proposed approach for out-of-distribution data. For this, especially the relevance of using an ensemble of models is expected to be different.

6. References

[1] A. Sansavini, M. E. Favilla, M. T. Guasti *et al.*, “Developmental Language Disorder: Early Predictors, Age for the Diagnosis, and Diagnostic Tools. A Scoping Review,” *Brain Sciences*, vol. 11, no. 5, p. 654, 2021.

[2] X. Wu, K. M. Knill, M. J. Gales, and A. Malinin, “Ensemble Approaches for Uncertainty in Spoken Language Assessment,” in *Proceedings INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020, pp. 3860–3864.

[3] T. Kemp and T. Schaaf, “Estimating confidence using word lattices,” in *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*. ISCA, 1997, pp. 827–830.

[4] G. Evermann and P. Woodland, “Large vocabulary decoding and confidence estimation using word posterior probabilities,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3. IEEE, 2000, pp. 1655–1658.

[5] A. Ali and S. Renals, “Word Error Rate Estimation Without ASR Output: E-WER2,” in *Proceedings INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020, pp. 616–620.

[6] A. Ragni, Q. Li, M. J. F. Gales, and Y. Wang, “Confidence Estimation and Deletion Prediction Using Bidirectional Recurrent Neural Networks,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 204–211.

[7] A. Malinin and M. Gales, “Uncertainty Estimation in Autoregressive Structured Prediction,” in *International Conference on Learning Representations (ICLR)*, 2021.

[8] D. Oneaţă, A. Caranica, A. Stan, and H. Cucu, “An Evaluation of Word-Level Confidence Estimation for End-to-End Automatic Speech Recognition,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 258–265.

[9] 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.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Curran Associates Inc., 2017, pp. 6405–6416.

[11] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.

[12] 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.

[13] 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.

[14] 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.

[15] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, “Unsupervised Domain Adaptation for Speech Recognition via Uncertainty Driven Self-Training,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6553–6557.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[17] 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.

[18] 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.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[20] M. Ravanelli, T. Parcollet, P. Plantinga *et al.*, “SpeechBrain: A General-Purpose Speech Toolkit,” 2021.

[21] A. Malinin, A. Ragni, K. Knill, and M. Gales, “Incorporating Uncertainty into Deep Learning for Spoken Language Assessment,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2017, pp. 45–50.