ZHAW-InIT at GermEval 2020 Task 4: Low-Resource Speech-to-Text

Matthias Büchi, Malgorzata Anna Ulasik, Manuela Hürlimann,
Fernando Benites, Pius von Däniken, and Mark Cieliebak

Institute of Applied Information Technology
Zurich University of Applied Sciences
{*buec, ulas, hueu, benf, vode, ciel*}*@zhaw.ch*

CEUR-WS.org/Vol-2624/germeval-task4-paper3.pdf

Abstract

This paper presents the contribution of ZHAW-InIT to Task 4 "Low-Resource STT" at GermEval 2020. The goal of the task is to develop a system for translating Swiss German dialect speech into Standard German text in the domain of parliamentary debates. Our approach is based on Jasper, a CNN Acoustic Model, which we fine-tune on the task data. We enhance the base system with an extended Language Model containing in-domain data and speed perturbation and run further experiments with post-processing. Our submission achieved first place with a final Word Error Rate of 40.29%.

1 Introduction

Automatic Speech Recognition (ASR) is defined as mapping audio signals to text. A particular challenge for ASR arises if a language does not have a standardized writing system, as is the case for Swiss German. In German-speaking Switzerland, Swiss German is the default spoken language on most occasions, from formal to informal; however, the language of reading and writing is Standard German ("medial diglossia", Siebenhaar and Wyler (1997)). Swiss German is increasingly used for writing in informal contexts, especially on social media, but users usually write phonetically in their local dialect (Siebenhaar, 2013). The particular dialects of Swiss German differ from each other to such an extent that speakers of one dialect might even have difficulty understanding dialects from some other regions. An indirect consequence is that many dialects are considered low-resource, since there is not enough data for each dialect for many natural language processing tasks. Nonetheless, there is enough data to train ASR systems for Standard German, which is spoken by a substantially larger group of native speakers, being an official language also in Germany and Austria. On official occasions, speeches are written down, transcribed, or logged in Standard German. Since the linguistic distance between the Swiss German dialects and the official language German are quite large, this poses a similar task as Cross-Linguistic Speech-To-Text (CL-STT; also referred to as speech-to-text translation) which is a difficult interdisciplinary challenge, combining STT with elements of Machine Translation (MT) (Bérard et al., 2016). Both fields have a long history of methods and approaches, which are currently at the point of converging thanks to the development of deep learning technology. This combination of ASR and MT is indeed needed in the context of Swiss German dialects, as speeches are paraphrased or even translated (see Section 6 for an example).

The Shared Task "Low-Resource STT" at GermEval 2020 aimed exactly at a specific Swiss case of CL-STT: translating Swiss German dialect spoken in an official context to written Standard German.

In our approach, we applied a general character-based ASR system (Li et al., 2019), pre-trained on a large German corpus, and fine-tuned to the Shared Task data. We further enriched our Language Model with additional publicly available data.

2 Shared Task Description

The goal of this Shared Task was to develop a system for translating Swiss German dialect speech into Standard German text in the domain of par-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

liamentary debates.

A data set of 36'572 utterances with a total duration of 69.8 hours was made available for training the systems and a 4 hour test set was used for evaluating solutions. The training data consists of a set of recordings of debates held in the parliament of the canton of Bern, with utterances produced by 191 speakers. None of these 191 speakers occur in the test set. The audio recordings contain mostly Swiss German dialect speech with the majority of the utterances being spoken in Bernese dialect; however, there are also some recordings of Standard German speech as well as a few English utterances. Each utterance contains one sentence and has an average duration of 6.9 seconds.

All recordings have been manually transcribed into Standard German, while the alignment between audio and transcripts was performed automatically by the task organizers (Plüss et al., 2020b,a).

The transcript accuracy is measured with the Word Error Rate (WER), which is the standard ASR evaluation metric. It is computed as the sum of the number of insertions, deletions and substitutions between predicted and reference sentences divided by the number of words in the reference (Zechner and Waibel, 2000). Selecting WER instead of the BLEU score, which is usually applied for automatic evaluation of translations, is justified by the task organizers with the fact that the Swiss German spoken in the parliament is comparatively close to Standard German and the diversity of the possible correct translations is very limited. Prior to evaluation, the task organizers normalized both ground truth and transcribed utterances by lowercasing them and removing punctuation.

3 Related Work

The most recent developments in both ASR and machine translation involve generalized methods that can be relatively easily ported across the two tasks, such as the encoder-decoder architecture. One of the most prominent, "Listen, Attend and Spell" (LAS) (Chan et al., 2016), uses an encoder-decoder architecture with attention and a pyramidal LSTM for the encoder. Chiu et al. (2018) describe improvements to LAS, such as multi-head attention, scheduled sampling, and label smoothing, which achieved new state-of-the-art performance, although only on proprietary voice search data. Other encoder-decoder models include the Neural Transducer (Jaitly et al., 2016), Recurrent Neural Aligner (Sak et al., 2017) and models based on the Transformer architecture (Vaswani et al., 2017) as in Dong et al. (2018). Zeghidour et al. (2018) achieved state-of-the-art performance on an end-to-end system based on convolutional neural networks (CNN). Their system can predict characters directly from raw waveforms, instead of the commonly used log-MEL features.

Li et al. (2019) propose a convolutional network with residual connections, with state-of-the-art results on the LibriSpeech and Wall Street Journal ASR data sets. The network predicts a character at each step (of 20 ms) and a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) is used for training. Beamsearch decoding allows the prediction to match a pre-trained Language Model. Preliminary work, Büchi (2020), showed that this approach was much easier to adapt and train on a large German corpus in comparison to hybrid systems trained with Kaldi (Povey et al., 2011) which achieve similar results.

While usually tens of thousands of hours of audio are required for achieving state-of-the-art ASR performance, some approaches target languages where only a few hours of data are available (Samarakoon et al., 2018). The use of pre-training and transfer learning are especially helpful in such challenging setups (Stoian et al., 2019).

Although there are approaches which directly target the speech translation setup (Bérard et al., 2016; Jia et al., 2019b,a), and the Shared Task data consists of translations and paraphrases of the spoken utterances, we decided not to add an additional component dealing specifically with translation to our system because of the lack of relevant available data.

4 System Description

This section describes the initial system used to establish a base for our experiments. Important concepts as well as parameters crucial for the experiments are explained.

4.1 Reference Text Pre-processing

We normalized all texts before training the Acoustic Models and Language Models. This step was necessary to have a standardized set of possible characters, which in this case were the letters a-z, ä, ö and ü. Normalization was performed in multiple steps, starting by lower-casing the whole

text and splitting it into sentences. All punctuation symbols were removed, except for points and commas which might be used as decimal point or for ordinal numbers. Numbers were transliterated to words. Common abbreviations and symbols were replaced by their spoken form (e.g. "%" by "Prozent" or "kg" by "Kilogramm"). Letters with diacritics other than ä, ö, and ü were replaced by their counterpart without diacritics. Finally, any remaining unknown symbols were removed without replacement.

4.2 Acoustic Model

An Acoustic Model was used to predict linguistic units based on an audio signal. For this purpose, Jasper (Li et al., 2019), a DNN-based model, was applied. Jasper predicts a probability distribution over all possible characters at every time step based on mel-filterbank features as input. The input was augmented with SpecAugment (Park et al., 2019).

The model consists of convolutional layers structured in blocks and sub-blocks. A model $B \times R$ is defined by the number of blocks B and number of sub-blocks R. Every sub-block consists of a 1D-convolution, batch-normalization, ReLU, and dropout. The input of each block is connected to the last sub-block by a residual connection. We applied the Dense Residual configuration, which is shown in Figure 1, where the output of each block is additionally added to the inputs of all following blocks. For pre- and post-processing one and three additional blocks were used, respectively.

During training, the CTC loss (Graves et al., 2006) was minimized using the Novograd optimizer introduced in Li et al. (2019).

4.3 Decoding

In order to get transcriptions from the Acoustic Model output, beam search was applied. Beam search tries to find the most probable text sequence given probabilities of characters over time. Additionally, a Language Model was used to re-rank the beam search hypotheses. A Language Model penalizes words that are not known and assigns a probability to each word given the words preceding it. The weight of the Language Model is controlled with parameter α. A parameter β is used as the word insertion bonus to prevent the preference of long words. The Language Model we used was a 6-gram model trained with KenLM (Heafield, 2011).

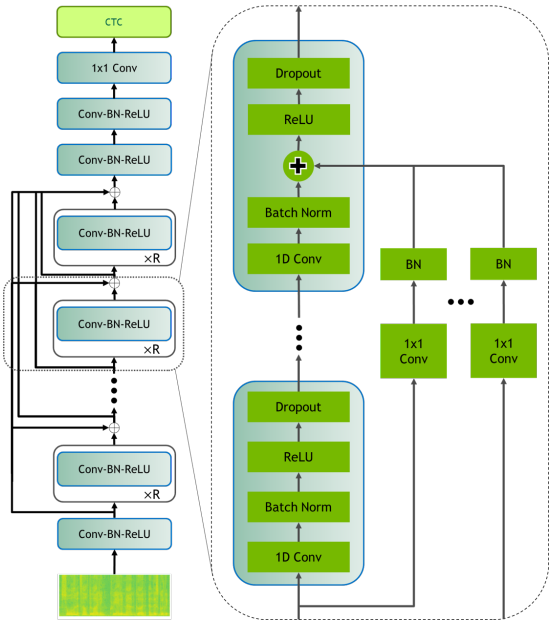

Figure 1: Visualization of the Jasper $B \times R$ Dense Residual model, from the Jasper Github repository (NVIDIA, 2020). It shows one pre-processing, three post-processing and intermediate blocks with residual connections.

4.4 Pre-training on Standard German

The Acoustic Model requires a large amount of data for training. Therefore, Standard German speech data as listed in Table 1 was used to create a pre-trained model[1]. Based on the given data sets, a combined version was created. Training, development and test splits were kept if given in the original data sets. Otherwise, custom splits were created with a size of 15% for test and validation, but with a maximum of 15000 seconds.

For the size of the model the configuration 10×5 was used. The model was trained with an initial learning rate of 0.015 on batches of size 64 for a total of 100 epochs.

4.5 Fine-Tuning

The pre-trained model was used as a base for fine-tuning using the task specific data. The first few blocks serve as acoustic feature extraction. Since acoustic features of Standard German and Swiss German are very close, only weights of the post-processing blocks as well as the last three or five intermediate blocks were updated, depending on

[1] Accessible through https://github.com/german-asr/megs.

Table 1: List of speech corpora used for pre-training. We used the original training splits, if available, and removed all identified invalid samples (e.g. containing wrong transcriptions or corrupted audios). This resulted in training data consisting of 536.9 hours of speech.

Name	Size (h)	Num. of Speakers
TuDa (Milde and Köhn, 2018)	183	179
CV (Ardila et al., 2019)	324	4852
VoxForge (VoxForge, 2019)	32	328
SWC (Baumann et al., 2018)	266	573
M-AILABS (M-AILABS, 2019)	233	-

the experiment as described in Section 5.2. Apart from the frozen blocks, the same hyperparameters were used as for the pre-training. The model was trained for another 100 epochs for fine-tuning (see Figure 2 for Word Error Rate progression over the 100 epochs).

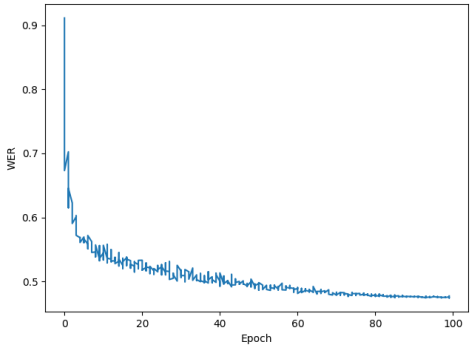

Figure 2: Word Error Rate progression on the internal development set.

4.6 Performance

The acoustic models were trained on a NVIDIA DGX-1 system. Pre-training with about 540 hours of Standard German took approximately 197 hours using two NVIDA Tesla V100 GPUs, while fine-tuning of the acoustic model (AM-A-5x5-SP) with about 70 hours of Swiss German speech required approximately 21 hours with one V100 GPU. The time for inference was much lower and took only about two minutes per 4 hours of speech on a NVIDA Titan X GPU. Applying the language model (LM extended) required some additional computation time. However, this took only a few minutes on a recent system for training as well as for decoding in combination with the beam search algorithm.

5 Experiments

We describe the experiments we conducted in order to improve the baseline system in Section 5.2, present the results we obtained in Section 5.3 and reflect on them in 5.4.

5.1 System Components

The data set provided as part of the Shared Task was split into internal train, development and test sets. The train set consisted of 32'978 utterances, the development set contained 1'778 utterances, while the test set comprised 1'816 utterances. This split approximates 90% training, 5% development, 5% testing. A single speaker could not occur in different sets and the utterance lengths were taken into account for splitting.

The experiments consisted in fine-tuning the baseline system with the use of additional text data and, in one case, in applying transcript post-processing.

Acoustic Models The baseline Acoustic Model (called "AM base" below) was fine-tuned on the internal train set, first on three blocks (model "AM-E 3x5") and in the second version on five blocks (model "AM-E 5x5"). In the last step of Acoustic Model fine-tuning, the baseline model was re-trained on the complete official train set (internal train, development and test sets combined), which resulted in the model called "AM-A 5x5". Additionally, we trained a model with the internal training set without applying any pre-training (model "AM-NOPRE").

Language Models The language modelling setup is described in Section 4.3. We used two different Language Models (LMs). The basic Language Model ("LM base") consists of corpora 1-3 in Table 2. Since these corpora are from different domains than the task data, we injected additional data to fine-tune

Table 2: List of text corpora used for training Language Models. The first three corpora were used for the basic Language Model, while the last two were additionally included in the extended LM.

	Name	Num. of Sentences
1	News-Commentary (Bojar et al., 2018)	383‘764
2	EuroParl (Koehn, 2005)	1‘920‘208
3	Tuda-Text (Milde and Köhn, 2018)	7‘776‘674
4	Federal Chancellery Press Releases	174‘520
5	Training set transcripts	32‘977

the LM: corpus 4 is a collection of 11’576 press releases by the Federal Chancellery (Bundeskanzlei). These were scraped from `https://www.bk.admin.ch/bk/de/home/dokumentation/medienmitteilungen.msg-id-<ID>.html` using a custom script, where consecutive <ID>s up to the most recent press release were queried and the content was subsequently extracted using XPath. Corpus 5 consists of the internal training set transcripts. The LM trained on all available corpora (1-5) is referred to as ”LM extended”.

Article Post-processing During development we noticed that there was a considerable amount of errors due to incorrectly predicted articles (e.g ”der”, ”die”, ”das”) (see Section 5.4 for more details). We identified individual definite and indefinite articles in a predicted utterance, removed them, and queried the top 5 predictions of a BERT model (Devlin et al., 2019). If the originally predicted article appeared in the list of suggestions, we kept it. Otherwise it was replaced by the article scored highest by BERT, making sure not to replace an indefinite article by a definite one or vice-versa.

5.2 Experimental Setup

In total, nine experiments were conducted with the goal to investigate system performance of the various models. The details of the experiments are presented in Table 3. The very first experiment (”base”) was performed without any fine-tuning or post-processing on the base model, while the second one (”AMext3x5”) aimed at evaluating the predictions from the ”AM extended 3x5” model without applying any Language Model. In the third experiment we evaluated the model trained only on the internal Swiss German train set without any pre-training on Standard German (”AMno_pretrain”). The next two experiments consisted in introducing and extending the Language Model (”AMch3x5_LMbase” and ”AMch3x5_LMext”). Following that, we investigated data augmentation possibilities. In addition to SpecAugment which is used in all experiments, we applied speed perturbation (Ko et al., 2015) on the Acoustic Model data (model ”AMch3x5_sp_LMext”). The sixth experiment (”AMch3x5_sp_LMext_artc”) was an attempt to improve the results by performing transcript post-processing. We sought to reduce the number of substitutions resulting from incorrect prediction of articles by applying BERT as described above. In ”AMch5x5_sp_LMext” we introduced the Acoustic Model ”AM extended 5x5” and replaced it with ”AM all 5x5” in the final experiment (AMall5x5_sp_LMext).

5.3 Evaluation

The results of all experiments were evaluated on the internal test set, except for the last one, ”AMall5x5_sp_LMext”, where the internal test set was used for training the models. The five best-performing versions were submitted for evaluation on the public test set of the Shared Task. Table 3 provides an overview of all results.

Eventually, we achieved 40.29% WER on the official test set. Our best performing system is a combination of the baseline Acoustic Model re-trained on 5 blocks with Swiss German data, speed perturbation, and a Language Model fine-tuned on in-domain data from Switzerland.

5.4 Discussion

The two largest performance improvements were achieved by fine-tuning the Acoustic Model on the task-specific data (”AMext3x5” vs ”base”: WER reduced by 38% absolute) and by using a general-purpose Language Model during decoding (”AMext3x5_LMbase” vs ”AMext3x5”: WER reduced by 7.64% absolute). Both of these are standard practices in ASR and hence these improvements are neither surprising nor particularly

Table 3: Experiments overview. Note on Acoustic Models: AM-E-3x5-SP stands for AM extended 3x5 with speed perturbation, and AM-A-5x5-SP for AM all 5x5 with speed perturbation.

System Name	Acoustic Model	Language Model	Post-Processing	WER internal	WER official
base	AM base	-	-	92.1%	-
AMext3x5	AM-E 3x5	-	-	54.1%	-
AMch3x5_LMbase	AM-E 3x5	LM base	-	46.46%	-
AMno_pretrain_LMext	AM-NOPRE SP	LM extended	-	46.82%	43.52%
AMch3x5_LMext	AM-E 3x5	LM extended	-	45.52%	42.61%
AMch3x5_sp_LMext	AM-E 3x5 SP	LM extended	-	44.83%	41.76%
AMch3x5_sp_LMext_artc	AM-E 3x5 SP	LM extended	articles	45.17%	42.2%
AMch5x5_sp_LMext	AM-E 5x5 SP	LM extended	-	44.43%	41.16%
AMall5x5_sp_LMext	AM-A 5x5 SP	LM extended	-	-	40.29%

insightful.

We identified articles as one distinct source of errors: around one sixth of substitution errors were articles; hence, we decided to address these during post-processing (model "AMch3x5_sp_LMext_artc"). Our method using BERT (see Section 5.2) did not improve performance. There are several reasons for this. First, while some articles were indeed improved with this method, often there was insufficient context to accurately determine the correct article. Domain-specific abbreviations (e.g. party names such as SVP, EVP) also proved difficult. Second, we observed a number of article errors that are due to the non-exact nature of the transcription. These are linguistic or stylistic changes and improvements of the spoken text and can therefore not be addressed by our method. For example: changing a spoken definite article to an indefinite one, using plural instead of singular, transcribing a spoken "es" with "das", or inserting an extra article into a coordinated noun phrase.

Finally, there is also a challenge that relates to the specific language variety in this task: articles in Swiss German are rather difficult to detect as they usually consist of single phonemes which are assimilated to the following noun. This means that articles may be missed at an earlier stage of processing and will not be present in the output passed to the post-processing.

Our extended Language Model brought a nearly 1% absolute WER improvement ("AMch3x5_LMext" vs "AMch3x5_LMbase"), which is less than we expected. However, this can be explained by the rather small amount of additional data - corpora 4 and 5 (see Table 2) only account for 2% of all sentences passed to the LM. Using more in-domain data in the LM could lead to a larger effect.

Further small improvements were obtained by using speed perturbation ("AMch3x5_sp_LMext" vs "AMch3x5_LMext": -0.7% absolute on our internal test set and -0.85% on the task test set) and retraining five Jasper blocks instead of three ("AMch5x5_sp_LMext" vs "AMch3x5_sp_LMext": -0.4% absolute on our internal test set and -0.6% absolute on the task test set).

We also note that our performance on the task test set is consistently better than the one on our internal test set.

6 Training Data Challenges

Before we conclude, we would like to reflect on the properties of the task data and their repercussions for WER results.

Our analysis of the errors and the data showed that properties of the data often lead to an increase in WER, where the ASR model provides an adequate transcription but is "punished" by data artefacts. We identified the following main issues:

- We noticed that transcriptions in the training set are inconsistent with respect to numerals, which are written as either numbers or words, so that transcribing the numeral four as "vier" when the reference transcript has "4" will lead to a substitution error. Since there is no consistency in the writing of numerals (e.g. always using words, always using numbers, using words when smaller than ten, etc), this leads to errors that we could not prevent.

- Transcripts are polished (e.g. speech disfluencies such as repetitions, hesitations, and false starts are removed) and reformulated so they become more readable, which means they do not exactly represent the spoken text. For example, in training set item `19940.flac`, the speaker starts by saying "mer hie enne" (DE: "wir hier drin", EN: "we in here"), but this was transcribed as "wir in diesem Saal" (EN: "we in this chamber"), leading to three errors (two substitutions and one deletion) when transcribed faithfully to the spoken utterance by the model.

- We also note issues with the segmentation of audio files, which, according to the task organizers, was performed automatically. This leads to insertion errors (when extra audio is included beyond what is transcribed) or deletion errors (when portions of the audio are missing) of the model that cannot be mitigated.

Given the observed discrepancies between the speech and transcript, additional evaluation measures might be considered. In CL-STT, BLEU scores are used for evaluation. Even though this metric has been criticized, it might fit the setup of this task better, since the paraphrasing might not be unique. Further, measures considering semantics and synonyms (Wieting et al., 2019; Kane et al., 2020) might prove helpful. However, in this specific case of official transcriptions, this would entail re-annotation, the cost of which would be prohibitive.

7 Conclusion

In this paper, we presented our contribution to the Shared Task on Low-Resource STT at GermEval 2020. Our solution consists of a CNN acoustic model based on Jasper (Li et al., 2019) with beam-search decoding and CTC loss. Our most successful model uses Transfer Learning, where we retrain the last five blocks of the Acoustic Model on the task data. Additionally, we use speed perturbation and a Language Model trained on both out-of-domain and in-domain text data. These improvements reduced the WER by over 50% compared to the Standard German baseline system. Our best model achieved a WER of 40.29% on the official task test set, resulting in first place out of three contributions.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common Voice: A Massively-Multilingual Speech Corpus. *ArXiv*, abs/1912.06670.

Timo Baumann, Arne Köhn, and Felix Hennig. 2018. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. *arXiv preprint arXiv:1612.01744*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Matthias Büchi. 2020. Speech Recognition Component for Search-Oriented Conversational Artificial Intelligence. Master's thesis, ZHAW Zurich University of Applied Sciences.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-Transformer: a No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Navdeep Jaitly, Quoc V Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, and Samy Bengio. 2016. An Online Sequence-to-Sequence Model Using Partial Conditioning. In *Advances in Neural Information Processing Systems*, pages 5067–5075.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019a. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019b. Direct Speech-to-Speech Translation with a Sequence-to-Sequence model. *arXiv preprint arXiv:1904.06037*.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral Based Interchangeability Assessor for Text Generation.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio Augmentation for Speech Recognition. In *INTERSPEECH*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. 2019. Jasper: An End-to-End Convolutional Neural Acoustic Model. *arXiv preprint arXiv:1904.03288*.

M-AILABS. 2019. M-AILABS Speech Dataset. https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/. Accessed: 2019-12-10.

Benjamin Milde and Arne Köhn. 2018. Open Source Automatic Speech Recognition for German. In *Proceedings of ITG 2018*.

NVIDIA. 2020. Jasper source code. https://github.com/NVIDIA/DeepLearningExamples/blob/master/PyTorch/SpeechRecognition/Jasper/images/jasper_dense_residual.png. Accessed: 2020-05-14.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *ArXiv*, abs/1904.08779.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020a. Forced alignment of swiss german speech to standard german text. In preparation.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020b. Germeval 2020 task 4: Low-resource speech-to-text. In preparation.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel. 2011. The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Hasim Sak, Matt Shannon, Kanishka Rao, and Françoise Beaufays. 2017. Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping. In *Interspeech*, volume 8, pages 1298–1302.

Lahiru Samarakoon, Brian Mak, and Albert YS Lam. 2018. Domain Adaptation of End-to-end Speech Recognition in Low-Resource Settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 382–388. IEEE.

Beat Siebenhaar. 2013. Sprachgeographische aspekte der morphologie und verschriftung in schweizerdeutschen chats.

Beat Siebenhaar and Alfred Wyler. 1997. *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. Pro Helvetia.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2019. Analyzing ASR Pretraining for Low-Resource Speech-to-Text Translation. *arXiv preprint arXiv:1910.10762*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.

VoxForge. 2019. VoxForge. http://www.voxforge.org/de. Accessed: 2019-12-10.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training Neural Machine Translation with Semantic Similarity. *arXiv preprint arXiv:1909.06694*.

Klaus Zechner and Alex Waibel. 2000. Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Neil Zeghidour, Qiantong Xu, Vitaliy Liptchinsky, Nicolas Usunier, Gabriel Synnaeve, and Ronan Collobert. 2018. Fully Convolutional Speech Recognition. *arXiv preprint arXiv:1812.06864*.