# GermEval 2020 Task 4: Low-Resource Speech-to-Text

**Michel Plüss** **Lukas Neukom** **Manfred Vogel**
Institute for Data Science
University of Applied Sciences and Arts Northwestern Switzerland
Windisch, Switzerland
michel.pluess@fhnw.ch



## Abstract

We present the results and findings of GermEval 2020 Task 4 on Low-Resource Speech-to-Text. Participants were asked to build a system translating Swiss German speech to Standard German text and minimize its word error rate. The task was based on a new dataset for Swiss German to Standard German speech translation, which contains 74 hours of sentence-level speech-text-pairs. 3 teams participated, with the winning contribution reaching a word error rate of 40.29 %.

## 1 Introduction

Speech-to-text methods for well-resourced languages like English or Standard German work very well. Lüscher et al. (2019) set the current state-of-the-art on the popular LibriSpeech test-other benchmark (Panayotov et al., 2015) with a word error rate (WER) of as low as 5 %. This is in stark contrast to the situation for Swiss German, the most frequently used spoken language in Switzerland, for which almost no publicly available training data for speech-to-text is available. Apart from the comparatively low number of speakers of around 5 million, the main reason for this is the lack of a standardized writing system. Thus, the official written language in the German part of Switzerland is Standard German, not Swiss German. Despite this, a lot of Swiss German speakers write in Swiss German, especially in informal conversations using messaging apps, but they resort to phonetical writing in their local dialect. The multitude of different written variants for each word makes direct speech-to-text almost impossible. Therefore, for most use cases, a speech-to-text method for Swiss German has to simultaneously translate to Standard German. This combination of speech recognition and translation is also referred to as speech translation.

For this shared task, we built a new dataset for Swiss German to Standard German speech translation. We provide data for training containing audio and text as well as data for testing containing only the audio. The testing data has to be transcribed and submitted on the task website for evaluation.

The remainder of this paper is structured as follows: The task and the evaluation of submissions are described in section 2. Our speech translation dataset is introduced in section 3. An overview of the submissions and results of this task can be found in section 4. Finally, section 5 wraps up the paper and gives directions for future work.

## 2 Task Description and Evaluation

The goal of the task is to build a sentence level Swiss German to Standard German speech translation system. The submission with the lowest WER wins. We chose WER as opposed to the BLEU score (Papineni et al., 2002), which is often used for machine translation and speech translation, for 2 main reasons. Firstly, manual inspection of a sample of our dataset showed that in most cases, the Standard German transcription does not deviate much from the Swiss German speech. Secondly, we only have a single reference transcription per utterance but BLEU would require multiple reference transcriptions to work well. Additionally, there is not a lot of margin for alternative transcriptions.

Apart from working with the provided dataset for training, participants were encouraged to explore data augmentation methods and transfer

learning approaches to build a better model in this low-resource setting.

At evaluation time, the following pre-processing steps are taken before comparing a submission against the reference transcriptions:

- Transform to lower case
- Remove punctuation , ; : . ? !
- Remove leading and trailing whitespace

Numbers are not normalized in any way.

## 3 Data

For this task, we created the, to the best of our knowledge, first publicly available dataset[1] for Swiss German to Standard German speech translation. It consists of 70 (train) + 4 (test) hours of mostly Swiss German and some Standard German speech from the parliament of the canton of Bern and corresponding Standard German transcriptions. The Swiss German speech is predominantly in the Bernese dialect. Some parliament members speak in Standard German, hence the small part of Standard German speech. Speech-text-pairs consist of a single sentence and were attained from the raw data using a fully automated alignment procedure described in (Plüss et al., 2020). The raw data consisted of audio recordings of full meetings, usually between 2 and 3 hours, and the transcript in a PDF file[2]. While the alignment quality is fairly good, it is certainly not perfect. The most common errors are missing or additional words at the beginning or end of a speech utterance compared to the transcript. Table 1 lists a few examples of this. This obviously makes the task harder and leads to non-avoidable mistakes. The transcripts were pre-processed as follows:

- Transform to lower case
- Replace or remove all characters except a-z, ä, ö, ü, 0-9, space, punctuation , ; : . ? !
- Remove leading and trailing whitespace

Details of the replacement and removal operations can be found in our code on GitHub[3].

[1] https://drive.switch.ch/index.php/s/PpUArRmN5Ba5C8J

[2] https://www.gr.be.ch/gr/de/index/sessionen/sessionen.html

[3] https://github.com/festivalhopper/germeval-2020-task-4/blob/master/transcript_preprocessing.py

The structure of the data is the same as in the Mozilla Common Voice project[4]. Table 2 gives a short description of the metadata provided in the TSV file.

## 4 Results

3 teams participated in the shared task and submitted a solution. Table 3 shows an overview of the results on the public part of the test set.

The team in first place, Büchi et al. (2020), achieved a WER of 40.29 %. Their approach is based on a CNN acoustic model called Jasper (Li et al., 2019). It is first trained on additional Standard German data and then fine-tuned on the task data. The model was trained using the CTC loss function (Graves et al., 2006). To further improve the results, a language model and data augmentation methods were applied. Table 4 shows some examples of true and predicted sentences and the corresponding WER. Sentence 1 is a good prediction, especially considering that the words ”gefährdet ist” are missing in the recording due to an alignment error. In sentence 2, the alignment is perfect, but the model chooses the word ”Führungskontrolle” rather than “Feuerungskontrolleur”. This is a seldomly used word in Swiss German and therefore hard to get right. Sentence 3 is actually a good prediction of what can be heard in the recording, but the sentence was reformulated in the transcription. In this case, the BLEU score with multiple reference transcriptions would better fit the task. Finally, the predicted sentence number 4 does not make too much sense and shows that the model still has considerable potential for improvement.

The team in second place, Kew et al. (2020), achieved a WER of 45.45 %. They follow a DNN-HMM approach for the acoustic model using a time delay neural network. No additional speech-to-text data is used. They create a pronunciation lexicon specifically adapted to this task. Like Büchi et al., they use a language model and apply data augmentation methods.

The team in third place, Agarwal et al. (2020), achieved a WER of 58.93 %. Their approach is based on DeepSpeech (Hannun et al., 2014), an end-to-end deep learning system. They use cascaded transfer learning, first training the model with English data, then transferring to Standard German, then finally to Swiss German. The Archi-

[4] https://voice.mozilla.org/en/datasets

| Sentence in Recording | Sentence in Transcript |
|---|---|
| ...der Fall. Wir Motionäre wurden zusammen mit anderen Interessengruppen sehr schnell eingela... | Wir Motionäre wurden zusammen mit anderen Interessengruppen sehr schnell eingeladen. |
| Das Pricing des Stroms ist relativ klar, es ist geregelt und die Gewinnspanne garantiert. Ich will... | Das Pricing des Stroms ist relativ klar, es ist geregelt und die Gewinnspanne garantiert. |

Table 1: Examples of alignment errors.

| Attribute | Description |
|---|---|
| client_id | Speaker ID |
| path | Name of the audio file in the clips folder |
| sentence | Ground truth Standard German transcription |
| up_votes, down_votes, age, gender, accent | Not available in the current version of the dataset |

Table 2: Description of the metadata in the TSV file of the dataset.

| Rank | Team | WER in % |
|---|---|---|
| 1 | Büchi et al. | 40.29 |
| 2 | Kew et al. | 45.45 |
| 3 | Agarwal et al. | 58.93 |

Table 3: Overview of the shared task's results, taken from the public ranking on the 22nd of May 2020. The WER column shows the word error rate in % on the public 50 % of the test set.

Mob (Samardžić et al., 2016) dataset is used as additional Swiss German training data. Like the other participants, they use a language model and apply data augmentation methods.

## 5 Conclusion

We have described GermEval 2020 Task 4 on Low-Resource Speech-to-Text. The task used a newly created dataset for Swiss German to Standard German speech translation described in section 3. 3 teams participated in the task, with the winning team reaching a WER of 40.29 %. This is a good result given that few research has been done on this topic and considering the alignment errors apparent in the dataset due to the fully automated alignment procedure. An open question is how well this model would generalize to other Swiss German to Standard German speech translation datasets or to a Standard German speech-to-text task.

The evaluation of the results of all teams indicates that data augmentation methods and language models work well in this low-resource setting. More details about the individual systems can be found in their respective system description papers, which are published in the SwissText & KONVENS 2020 proceedings.

We have made the dataset publicly available[5] to the research community beyond the GermEval competition, hoping to facilitate future research on this important topic.

In future work, we plan to minimize errors made by the automatic alignment procedure and substantially increase the dataset size by aligning additional raw data.

## Acknowledgments

First of all, we would like to thank the parliamentary services of the canton of Bern for their work on the transcription of the debates and for publishing recordings and transcripts on their website. Without them, this task would not have been possible.

We would also like to thank the GermEval 2020 organizers for hosting the Low-Resource Speech-to-Text task and for replying promptly to all our inquiries.

We especially thank the GermEval 2020 Task 4 participants for their interest in the shared task, for their participation, and for their timely feedback, which have helped us make the shared task

[5] https://drive.switch.ch/index.php/s/PpUArRmN5Ba5C8J

| ID | True Sentence | Predicted Sentence | WER in % |
|---|---|---|---|
| 1 | insbesondere kann der kanton mit finanziellen zuschüssen steuernd eingreifen, wenn die versorgungssicherheit gefährdet ist. | insbesondere kann der kanton mit finanziellen zuschüsse steuernd eingreifen die versorgungssicherheit | 28.57 |
| 2 | der feuerungskontrolleur, der von den gemeinden gewählt und eingesetzt wird, ist neutral. | die führungskontrolle die von den gemeinden gewählt und eingesetzt wird ist neutral | 25.00 |
| 3 | dabei ist zu beachten, dass der sinn dieser brückenangebote auch von der mehrheit nicht infrage gestellt wird. | das ist vielleicht in die debatte auch wichtig dass der sinn dieser brückenangebote ist auch von der mehrheit nicht infrage gestellt | 52.94 |
| 4 | wissenschaft läuft nicht so, dass ein mäzen, wie ein hansjörg wyss und vor allem nicht er, mit seiner fachkompetenz , so in wissenschaftliche forschung reinreden würde. | es läuft nicht so dass denen ans jürg weise und von er mit seiner fachkompetenz einsprechen in wissenschaftliche forschung wissen | 52.00 |

Table 4: Examples of true sentences compared to the predicted sentences by Büchi et al. with the corresponding word error rate. Punctuation is removed before calculating the WER.

a success.

## References


Aashish Agarwal and Torsten Zesch. 2020. Ltl-ude at low-resource speech-to-text shared task: Investigating mozilla deepspeech in a low-resource setting. In preparation.

Matthias Büchi, Malgorzata Anna Ulasik, Manuela Hürlimann, Fernando Benites, Pius von Däniken, and Mark Cieliebak. 2020. Zhaw-init at germeval 2020 task 4: Low-resource speech-to-text. In preparation.

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*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Tannon Kew, Iuliia Nigmatulina, Lorenz Nagele, and Tanja Samardžić. 2020. Uzh tilt: A kaldi recipe for swiss german speech to standard german text. In preparation.

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.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. Rwth asr systems for librispeech: Hybrid vs attention - w/o data augmentation. In *INTERSPEECH*.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

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

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. Archimob - a corpus of spoken swiss german. In *Language Resources and Evaluation (LREC 2016)*, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 4061–4066. s.n.