

# Swiss German Speech to Standard German Text

## SwissText.org Shared Task 3

Alex Wolf  
University of Zurich  
alex.wolf@uzh.ch

Deborah Noemie Jakobi  
University of Zurich  
deborahnoemie.jakobi@uzh.ch

### Abstract

This paper presents a solution to the [Swiss-Text Conference \(STC\)](#) shared task 3, 2021. The shared task challenges participants to implement models for translating Swiss German speech to standard German text. The authors implemented multiple *DeepSpeech* models using the data provided by SwissText.org as well as additional data. The use of the additional data in combination with the official data lead to some promising improvements of the model trained on the official data only. An additional experiment with a sequence to sequence translation model was trained in order to improve our score. We achieved a BLEU score of up to 0.17 on the test set of SwissText. Although the performance of the final model did not score well in comparison with other submissions, there is plenty of room for possible improvements.

## 1 Introduction

Swiss German has a wide variety of different Swiss German dialects, with a huge difference in words, pronunciation, even to the point of sounding like a different language. Swiss German has relatively little speakers (around 5 million) and there is hardly any standardized spelling. This leads to Standard German being one of the official writing language in Switzerland. As there is no official Swiss German spelling, most speakers using written Swiss German just use their own spelling which resembles mostly a phonetical translation. This leads to huge variance within Swiss German writing ([Pluss et al., 2020](#)). It makes therefore sense to translate spoken Swiss German to Standard German in order to correspond to the official language situation. Tackling a standardized translation of different spoken Swiss German dialects into standardized German text requires a vast amount of data and fine tuning. The [STC](#) 2021 proposed a shared task to tackle this problem and provided a dataset ([Swiss-Text Conference dataset \(STCD\)](#)) to train and fine

tune on. This paper shows what kind of experiments, data, and approaches the authors used to tackle this problem. The proposed task is very complex as it includes not only a [Speech-to-Text \(STT\)](#) conversion but also translation from Swiss German to Standard German which can be referred to as speech translation ([Pluss et al., 2020](#)). Additionally, it possibly includes domain shift, as the training data stems only from Swiss parliament speeches while the domain of the test set is unclear. This paper presents a quick overview on the previous research done within this area, an introduction to the DeepSpeech model used for tackle the task and our experiments and results.

## 2 Literature review

add more general stuff about [STT](#).

The shared task of a translating spoken Swiss German into standard written German was already presented by the [STC](#) in 2020. While the current 2021 task is about reaching the highest BLEU score, last year's submissions were ranked based on the least [Word Error Rate \(WER\)](#). [Buchi et al. \(2020\)](#) achieved the best [WER](#) of 40.29%. The authors used a CNN acoustic model named Jasper. They used additional Standard German data and fine-tuned on the official data set. They used different augmentation techniques and a language model. [Pluss et al. \(2020\)](#) add text. ([Kew et al., 2020](#)) achieved the second best [WER](#). ...

? used an end-to-end model called DeepSpeech and achieved an [WER](#) of 58.93%. This is the model we decided to use as well. It is publicly available on GitHub and can easily be fine-tuned or used for transfer learning.

## 3 DeepSpeech

explain how deepspeech works...

## 4 Experiments

### 4.1 Training Data

To train our model we use a subset of the dataset provided by the [STC](#) as well as the ArchiMob corpus. The [STCD](#) contains 38 GB of labeled, 65 GB of unlabeled spoken swiss german audio data and an additional validation set containing 1.5 GB of data ([Ando and Zhang, 2005](#)). The ArchiMob corpus (Release 2) contains X GB of spoken swiss german data ([Samardžić et al., 2016](#)). We remove all audio files, from the [STCD](#), where the quality of the translation is rated lower than 0.7 (rating provided by [STC](#)). Additionally, we also need to remove files larger than XY GB due to memory limitations.

### 4.2 Evaluation metrics

We evaluate our model through two types of metrics. The BLEU score ([Papineni et al., 2002](#)), which is required by the [STC](#) to compare the results. The [STC](#) specifically requires a corpus based BLEU score (todo cite stc), which aims at measuring the distance/similarity of the generated text and the provided ground truth.

Additionally, we measure the [WER](#) for our models, as this is a common metric for speech recognition systems ([Park et al., 2008](#)).

### 4.3 Environment or Setup?

We train the model on a single XYZ GPU (XYG memory) for X epochs, with batch size of X and used a learning rate of XY, ....

We perform minimal tuning of our model's hyperparameters following the work of [Pluss et al. \(2020\)](#) as this model achieved the best results in previous challenges.

### 4.4 Models

Following [Pluss et al. \(2020\)](#) we use a DeepSpeech architecture ([Hannun et al., 2014](#)) as our main model for speech-to-text translation. In order to get better results we use a pre-trained DeepSpeech model ([Ohme, 2020](#)) as the base model for most of our experiments. The first model we trained served as our private baseline.

**Baseline model** We trained a bare DeepSpeech model, with the default hyperparameters from [Agarwal and Zesch \(2019\)](#), on the labelled [STCD](#), which achieved a BLEU score of XY BLEU. The

baseline model will be referred to as baseline model or model #0 henceforth.

**Pre-trained model #1** In order to improve our baseline model we use a pre-trained german DeepSpeech model ([Ohme, 2020](#)), which is fine-tuned on the [STCD](#) dataset for 75 epochs, with a learning rate of XY, and a batch size of XY. The model achieved a BLEU score of XY on the validation set and XY on the test set.

**Additional data model #2** To further improve our 1<sup>st</sup> model we added additional data to our fine-tuned model. We added the ArchiMob data and fine-tuned for an additional 75 epochs, with the same hyperparameters as in model #1 and achieved the following BLEU scores on the validation and training set, XYZ, XYZ2, respectively.

**ArchiMob data model #3** As the 2<sup>nd</sup> model did not achieve the expected results we decided to fine-tune our model on the ArchiMob data, as we expected the data distribution to be closer to the actual test set. The model was trained for 75 epochs, with a learning rate of XY, and a batch size of XY. We achieved the following BLEU scores for the validation and test set, XYZ, XYZ2, respectively.

**Augmented data model #4** In order to produce a more robust model, that can handle different pronunciations and speeds, we added data augmentation to our 3<sup>rd</sup> model. We fine-tuned our model with the following data augmentations:

- warp: "Applies a non-linear image warp to the spectrogram. This is achieved by randomly shifting a grid of equally distributed warp points along time and frequency axis" ([DeepSpeech, 2020](#))
- tempo: "Scales spectrogram on time axis and thus changes playback tempo." ([DeepSpeech, 2020](#))

The [STCD](#) dataset is used for the augmentation fine-tuning on the 3<sup>rd</sup> model. This approach achieved a BLEU score of XY on the validation set and XY on the test set. The model was trained for X epochs with a warp probability of 0.1 and a tempo probability of 0.1. We increased the batch size to X, as the augmentation process per batch took more time than expected. Due to time limitations we were not able to continue training the model any further, but the model was still improving and

we expect that this approach could produce even better results.

**Text-to-Text model #5** Our 5<sup>th</sup> approach tries to improve the output by adding a text-to-text translation model after the speech-to-text model. We fine tuned a pre-trained german to english model following [Tiedemann and Thottingal \(2020\)](#) using the "huggingface" framework ([Wolf et al., 2020](#)) on the [STCD](#) dataset. With this approach we achieved a BLEU score of XY on the validation dataset and XY on the test dataset.

## 5 Results & Discussion

Interestingly, we noticed that our internal deviations in the BLEU score are not as great as on the official test set.

Model#	Data	Train BLEU	Test BLEU
1	SwissText	0.23	0.0004
2	ArchiMob	0.27	<b>0.17</b>
3	ArchiMob	0.24	0.07

Table 1: Font guide.

## 6 Conclusion

### References

- Aashish Agarwal and Torsten Zesch. 2019. German end-to-end speech recognition based on deepspeech. In *KONVENS*.
- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.
- Matthias Buchi, Malgorzata Anna Ulasik, Manuela Hürlimann, Fernando Benites, Pius von Daniken, and Mark Cieliebak. 2020. Zhaw-init at germeval 2020 task 4: Low-resource speech-to-text.
- DeepSpeech. 2020. [Training your own model — deepspeech 0.9.3 documentation](#).
- Awni Y. Hannun, Carl Case, J. Casper, Bryan Catanzaro, G. Diamos, Erich Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *ArXiv*, abs/1412.5567.
- Tannon Kew, Iuliia Nigmatulina, Lorenz Nagele, and Tanja Samardžić. 2020. Uzh tilt: A kaldi recipe for swiss german speech to standard german text.
- Karsten Ohme. 2020. Pre-trained deepspeech model v0.9.0.
- Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Youngja Park, Siddharth Patwardhan, K. Visweswariah, and S. C. Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*.
- Michel Pluss, Lukas Neukom, and Manfred Vogel. 2020. Germeval 2020 task 4: Low-resource speech-to-text.
- Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. [ArchiMob - a corpus of spoken Swiss German](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.