ZHAW-CAI: Ensemble Method for Swiss German Speech to Standard German Text

Malgorzata Anna Ulasik, Manuela Hürlimann, Bogumila Dubel, Yves Kaufmann,
Silas Rudolf, Jan Deriu, Katsiaryna Mlynchyk, Hans-Peter Hutter, and Mark Cieliebak

Centre for Artificial Intelligence
Zurich University of Applied Sciences
{*ulas, hueu, deri, mlyn, huhp, ciel*}*@zhaw.ch*
bodubel@gmail.com, y.kaufmann@yagan.ch, silasrudolf@gmail.com

Abstract

This paper presents the contribution of ZHAW-CAI to the Shared Task "Swiss German Speech to Standard German Text" at the SwissText 2021 conference. Our approach combines three models based on the Fairseq, Jasper and Wav2vec architectures trained on multilingual, German and Swiss German data. We applied an ensembling algorithm on the predictions of the three models in order to retrieve the most reliable candidate out of the provided translations for each spoken utterance. With the ensembling output, we achieved a BLEU score of 39.39 on the private test set, which gave us the third place out of four contributors in the competition.

1 Introduction

Speech-to-Text (STT) enables transcribing spoken utterances into text. For successfully performing a transformation from speech to a text, the existence of a standardised writing system of the target language is of prime importance. This is where Swiss German [1] poses a substantial challenge: it does not have a standardised orthography since it functions as the default spoken language in both formal and informal situations, while for writing, the Standard German language is used. This phenomenon, called "medial diglossia" (Siebenhaar and Wyler, 1997), occurs in the entire German-speaking part of Switzerland, which is additionally characterised by a high dialect diversity. Swiss German is increasingly used for writing in informal contexts, but since there is no single standard writing system, Swiss German speakers usually write phonetically in their local dialect in informal situations (Siebenhaar, 2003). On formal occasions such as work meetings and political debate, speech is typically transcribed into Standard German. As there is a considerable linguistic distance between Swiss German dialects and Standard German, developing a model for transcribing Swiss German speech into Standard German text actually involves Speech Translation, which combines STT with Machine Translation (MT) (Bérard et al., 2016).

As a response to the Shared Task "Swiss German Speech to Standard German Text" organised at Swisstext 2021, we provided a solution consisting of three models based on different architectures: Fairseq (Wang et al., 2020a), Jasper (Li et al., 2019) and Wav2vec XLSR-5 (Baevski et al., 2020) which were trained with various data sets, both in Standard German and Swiss German. Their predictions were subsequently fed into a majority voting algorithm with the aim to select the most reliable translation.

The remainder of this paper is structured as follows: Section 2 provides the description of the Shared Task and Section 3 discusses relevant literature. In Section 4 we present the systems which make up our final solution, their architecture and the applied training data. In section 5 we provide an overview of all experiments performed with these models and their outputs. Section 6 lays out the ensembling approach and section 7 presents the post-processing experiments we performed on the predictions of the models. The paper ends with a conclusion presented in section 8.

2 Shared Task Description

The goal of the Shared Task was to build a system for translating speech in any Swiss German dialect into Standard German text (Plüss et al., 2021).

The organisers provided a labelled data set con-

[1] To be precise, there is no single "Swiss German" language, but rather a collection of many different regional dialects that are subsumed with this term.

taining 293 hours of audio recordings, mostly in the Bernese dialect, transcribed in Standard German. Since the alignment between the recordings and the transcripts was done automatically, each utterance has an Intersection over Union (IoU) score reflecting its alignment quality. Additionally, there was an unlabelled data set consisting of 1208 hours of recordings, mostly in the Zurich dialect. The solutions were evaluated based on a 13 hours test set, which contains recordings of speakers coming from all German-speaking parts of Switzerland. The dialect distribution of the test set is close to the actual Swiss German dialect distribution in Switzerland.

The translation accuracy of the provided solutions is measured using BLEU, a standard metric for automatic evaluation of machine translation (Papineni et al., 2002). The approach consists in counting n-grams in the candidate translation matching n-grams in the reference translation without taking the word order into account. The metric ranges from 0 to 100. A perfect match results in a score of 100. A score of 0 occurs if there are no matches. The tool used by the organisers for evaluating solutions is the NLTK implementation of the BLEU score with default parameters[2]. Prior to evaluation, both the references and the translations are normalised: the utterances are lowercased, the punctuation is removed, the numbers are spelled out and all non-ASCII characters except for the letters "ä", "ö", "ü" are removed.

The test set was split into a public and a private subset of equal sizes. For all evaluations presented in this paper, the public test set was used.

3 Related Work

Speech Translation (ST) is the task of translating spoken text in a source language to text or speech in a target language. The approaches to solve this problem can be put into two categories: cascading approaches and end-to-end approaches (Sperber and Paulik, 2020).

Cascaded Approaches work by splitting the task into two steps: first, an STT model transcribes speech of the source language to text in the target language, and then a machine translation (MT) module translates the generated text into the target language (Waibel et al., 1991). The main issue with the cascaded approach is the fact that errors made by the STT module are propagated to the MT module (Ney, 1999). Thus, efforts are put into coupling the STT and MT modules to prevent error propagation, for instance, by generating multiple hypotheses of the STT system via n-best search or the creation of lattices (Woszczyna et al., 1993; Schultz et al., 2004).

End-to-End Approaches model ST as a single task, where input is speech in the source language, and the output consists of text or speech in the target language. The main issue with this modelling approach is the lack of sufficient training data. Whereas data for STT typically consists of several hundreds of hours of transcribed data, most ST datasets contain only a fraction of this amount. For instance, the Europarl-ST corpus contains on average only 42 hours of transcribed data per language pair (Iranzo-Sánchez et al., 2020), whereas the Librispeech STT corpus contains around 1000 hours of transcribed data (Panayotov et al., 2015). For this reason, end-to-end approaches nowadays rely on leveraging multi-task learning and single language pre-training of the STT and MT submodules and use the ST dataset for fine-tuning (Wang et al., 2020b).

Most cascading approaches rely on data where access to both the source language transcript and its target language translation is needed. However, in our scenario, we do not have access to written text of the source language since Swiss German is a spoken language, and thus, often directly transcribed into Standard German (see 1 for more details). Thus, our models follow the End-to-End approach.

4 Systems Description

This section describes the architecture of the three models which build the foundation for the experiments presented in Section 5 and are components of the final solution which combines the three models' outputs in an ensembling algorithm. The section also explains what data was used for training the models.

4.1 Fairseq

4.1.1 Model

Fairseq is based on the transformer architecture for Speech-to-Text provided by Fairseq S2T Toolkit (Wang et al., 2020a), which combines the tasks of STT and ST under the same encoder-decoder

[2] https://www.nltk.org/api/nltk.translate.html#nltk.translate.bleu_score.corpus_bleu

architecture (Changhan Wang, 2020). The experiments were trained with the small transformer model with 256 dimensions, 12 Layers encoder, 6 Layers decoder, 27M parameters, Adam optimiser, and inverse square root for the learning rate scheduler. Decoding is executed with a character-based SentencePiece model (Taku Kudo, 2018) using an n-best decoding strategy with n=5. The acoustic model (encoder) can be pre-trained with the same transformer architecture as described above.

4.1.2 Data

The audios were extracted to 80-dimensional log mel-scale filterbank features (windows with 25 ms size and 10 ms shift) and saved in NumPy format for the training. To alleviate overfitting, speech data transforms SpecAugment (Park et al., 2019), adopted by Fairseq S2T, were applied. For text normalisation we used the script provided by the task organisers. Additional numbers were spelled out using num2words[3]. We use three additional datasets:

- SwissDial (Pelin Dogan-Schönberger, 2021): 26 hours of Swiss German
- ArchiMob (Tanja Samardzic, 2016): 80 hours of Swiss German
- Common Voice German v4: 483 hours of German[4]

The SwissDial dataset consists of 26 hours of audios in 8 different Swiss dialects with corresponding transcriptions in Swiss dialect and Standard German translations. The Swiss German transcription rules differ between dialects. ArchiMob contains 70 hours of audios in 14 different Swiss dialects with transcription in Swiss German, where each word is additionally provided with a Standard German normalisation. The transcription rules are normalised and are equal for all dialects (Dieth transcription, (Dieth and Schmid-Cadalbert, 1986)). Common Voice German v4 consists of 483 hours of audios in Standard German with corresponding transcriptions.

4.2 Jasper

4.2.1 Model

We used the Jasper (Li et al., 2019) configuration corresponding to our best submission in the predecessor of this Shared Task (Büchi et al., 2020). The Acoustic Model as per Büchi et al. (2020) consists of 10x5 blocks and was pre-trained on 537 hours of Standard German data (see Büchi et al. (2020), Table 2). In all reported experiments, we fine-tuned five blocks on the Shared Task data as described in Section 5.2 below. We used last year's extended language model, a 6-gram model trained with KenLM, without further fine-tuning on this year's data. For the data sources, see Table 2 in Büchi et al. (2020). Decoding was done using beam search with a beam size of 1024.

4.2.2 Data

We extracted the audios to 64-dimensional mel-filterbank features with 20ms window size and 10ms overlap as input to the Jasper acoustic model. The reference texts were preprocessed as described in Büchi et al. (2020). No additional Swiss German audio data was used for training Jasper.

4.3 Wav2vec XLSR-53

4.3.1 Model

Wav2vec XLSR-53 is a cross-lingual extension of wav2vec 2.0 as per Baevski et al. (2020). Pretrained on 53 different languages, it attempts to learn a quantisation of the latent representations shared across languages by solving a contrastive task over masked speech representations. In the experiment below, we fine-tuned wav2vec XLSR-53 on the Shared Task data. No explicit language model was used to conduct the experiment.

4.3.2 Data

The labelled data used for fine-tuning XLSR-53 was based on the task training data. However, it was further pre-processed removing all utterances which contained special characters or were detected as not being in German using langdetect[5]. Numeric values were replaced by strings using num2words[6].

5 Experiments on Individual Models

Sections 5.1 through 5.3 present the experiments we performed to improve the individual models and provide the BLEU scores achieved in each experiment. We also discuss approaches to improve the model outputs with the use of ensembling (Section 6) and post-processing (Section 7).

[3] https://pypi.org/project/num2words/
[4] https://commonvoice.mozilla.org/en/datasets/
[5] https://github.com/Mimino666/langdetect
[6] https://pypi.org/project/num2words/

5.1 Fairseq

Below we describe the different models and experimental results obtained with Fairseq. All Experiments are trained with the same configuration as described in Section 4.1 and can be divided into three groups: extension of training data, inclusion of a pre-trained encoder and ensembling.

5.1.1 Extending the training data

Fairseq F-SP-0.9 For `F-SP-0.9` we trained the model from scratch on the Shared Task training data. We used 176 hours, corresponding to an Intersection over Union (IoU) greater or equal to 0.9.

Fairseq F-SP-All We noted that the model `F-SP-0.9` generalises very poorly, so for `F-SP-All` we trained a new model with the entire task training data, which corresponds to 293 hours. Despite partially poorly aligned translations, the model benefits from the new data: the BLEU score is improved by about 4.32 points.

Fairseq F-SP-SD We decided to extend the training data with the SwissDial Corpus. For this, we trained a new model `F-SP-SD` with the entire task training data plus all data from SwissDial. This data extension improves the score by an additional 4.81 BLEU points in comparison to `F-SP-All`.

5.1.2 Including pre-trained encoder

Fairseq F-SP-DE We also investigated how to improve the encoder (acoustic model). We pre-trained a Standard German (DE) encoder on the Common Voice German v4 dataset. For `F-SP-DE`, we added the pre-trained encoder and trained the model on the entire Shared Task training data. Including the DE encoder improves the score by 3.36 BLEU points in comparison to `F-SP-All`.

Fairseq F-SP-SD-DE Since both models `F-SP-SD` and `F-SP-DE` improved the BLEU score, we decided to bring the two approaches together. We trained a new model `F-SP-SD-DE` with the entire Shared Task training data, SwissDial data and include the pre-trained DE encoder in the training. This brings an improvement of 8.37 BLEU points in comparison to `F-SP-All`.

Fairseq F-SP-AM-DE In this model we used the entire task training data plus the data from ArchiMob. For the training we included the pre-trained DE encoder. This setup improves the BLEU score by 14.01 in comparison to `F-SP-All`.

Fairseq F-SP-SD-CH In order to further improve the acoustic model, we trained an encoder in Swiss German (CH) on the SwissDial and ArchiMob dataset. We trained a new model `F-SP-SD-CH` with the entire Shared Task training data and SwissDial and included the pre-trained CH encoder in the training. The BLEU score in comparison to `F-SP-All` is improved by 12.54 points.

5.1.3 Ensembling

Fairseq Ensemble F-SP-SD & F-SP-DE (F-E1) In this experiment, we ensembled the models `F-SP-SD` and `F-SP-DE`. `F-E1` achieves a BLEU score of 28.74 . Ensembling is done with the implementation provided by the Fairseq S2T Toolkit[7]. In comparison to `F-SP-SD-DE`, which combines in the training setup the same training dataset SwissDial as `F-SP-SD` and the same DE encoder as `F-SP-DE`, the ensembling performs slightly better. In comparison to `F-SP-All` the BLEU score improves by 9.94 points.

Fairseq Ensemble F-SP-AM-DE & F-SP-SD-CH (F-E2) After the good performance of `F-E1`, we decided to ensemble `F-SP-AM-DE` and `F-SP-SD-CH`. This ensembling improves the BLEU score in comparison to `F-SP-All` by 17.00 points.

Fairseq F-E2 extended (F-E3) Finally, we trained a model on the entire available data for Swiss German (task, SwissDial and ArchiMob) and used this model to perform ensembling on top of `F-E2`. For time reasons, we were not able to complete the training and the output of this model could not been included in the final solution presented in 6. We only evaluated an intermediate status of the model and achieved a score of 36.83 BLEU points. In comparison to `F-SP-All`, it improves the score by 18.03 points.

Table 1 shows the public BLEU scores obtained with the Fairseq models on the Shared Task public part of the test set. The table contains additional information about applied train sets and encoders. `F-E3` achieved the best performance with a BLEU score of 36.83 on the public part of the test set (37.4 on the private part). In addition to ensembling, the inclusion of a CH encoder in

[7] https://github.com/pytorch/fairseq/issues/223

the training process as well as the extension of the training data with the ArchiMob corpus benefited the model performance most.

Table 1: Fairseq results.

Model	Train set	Encoder	BLEU
F-SP-0.9	task 0.9	training	14.48
F-SP-All	task all	training	18.8
F-SP-SD	task, SwissDial	training	23.61
F-SP-DE	task	DE	22.16
F-SP-SD-DE	task, SwissDial	DE	27.17
F-SP-AM-DE	task, ArchiMob	DE	32.81
F-SP-SD-CH	task, SwissDial	CH	31.34
F-E1	-	-	28.74
F-E2	-	-	35.80
F-E3	-	-	**36.83**

5.2 Jasper

Below we describe the different models and experimental results obtained with Jasper.

Jasper-FT For `Jasper-FT` we fine-tune the pre-trained Standard German model on the Shared Task training data. We used 169 hours, sampled from the set with an IoU greater or equal to 0.9, which were augmented to 507 hours using 90% and 110% speed perturbation as in Büchi et al. (2020).

Jasper-PL We noted that the task test set differs acoustically from the training data since different dialects are present and the audio quality tends to be lower. This motivated the creation of `Jasper-PL`, where we used pseudo-labeling on the test set. More precisely, we used the hypotheses of `Jasper-FT` on the task test set to fine-tune `Jasper-FT` for 20 additional epochs.

Jasper-PL-E We decided to further work on the (comparatively) low-quality audio of the task test set and used the Dolby Media Enhance API v1.1[8] to create an "enhanced" version of the task test set. The Enhance API automatically improves the quality of audio files, e.g. by correcting the volume and reducing noise and hum. We then fine-tuned `Jasper-FT` on this data, this time using the hypotheses provided by `Jasper-PL` as labels since these achieve a higher BLEU score.

Table 2 shows the public BLEU scores obtained with the Jasper models on the two different test sets (`Jasper-PL-E` was only evaluated on the enhanced test set). The best-performing Jasper model is `Jasper-PL` with a BLEU score of 32.97 on the public part of the test set. Using the enhanced audio data does not confer any advantage on either prediction or pseudo-label fine-tuning compared to the as-is data. We can, however, see the benefit of rather naive pseudo-labelling in this setting where training and testing data are quite different. Future work could expand on the use of pseudo-labelling by using more advanced setups, such as confidence-based (Kahn et al., 2020) or iterative (Xu et al., 2020) pseudo-labelling.

Table 2: Jasper results.

Model	Test set	BLEU
Jasper-FT	task	30.8
Jasper-FT	enhanced	26.4
Jasper-PL	task	**32.97**
Jasper-PL	enhanced	31.92
Jasper-PL-E	enhanced	32.92

5.3 Wav2vec XLSR-53

Below we describe the model and experimental results obtained with wav2vec XLSR-53.

wav2vec XLSR-53 FT For `wav2vec XLSR-53 FT` we fine-tuned the pre-trained baseline (as published on HuggingFace[9]) on the Shared Task training data. We used 227 hours, corresponding to an IoU greater or equal than 0.8. The data was pre-processed as outlined in Section 4.3.2.

Table 3: wav2vec XLSR-53 result.

Model	Train set	BLEU
wav2vec XLSR-53 FT	task 0.8	30.39

6 Ensembling

Having trained and evaluated the three models described in Sections 4.1, 4.2 and 4.3, we performed experiments with two ensembling methods: majority voting and a hybrid technique combining majority voting with perplexity calculation. We used the outputs of the best-performing models of each of the three systems, aiming to select the

[8] https://dolby.io/developers/media-processing/api-reference/enhance

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

most reliable translation for each utterance from among them. The best-performing models were `F-E2` (BLEU score of 35.80[10]), `Jasper-PL` (BLEU score of 32.97) and `wav2vec XLSR-53 FT` (BLEU score of 30.4).

The models were first categorised based on their BLEU scores into a primary, first auxiliary and second auxiliary models. `F-E2` with the highest score was selected as the primary model, `Jasper-PL` with the second best score was set as the first auxiliary model and `wav2vec XLSR-53 FT` was used as the second auxiliary model.

In the first step, we aligned the hypotheses of the three models and extracted text passages where all three hypotheses agree, leaving only text excerpts where the hypotheses disagree.

Majority Voting (MV) The majority voting consisted in collecting votes for each text excerpt defined in the previous step: a particular hypothesis receives a vote for each word it has in common with any other hypothesis. The hypothesis with the most votes is chosen as the best candidate translation. If multiple hypotheses score the same, the output of the model categorised higher in the hierarchy (primary, first auxiliary, second auxiliary) is selected.

Hybrid Ensembling (HE) The hybrid ensembling method combines majority voting with perplexity calculation. If more than one hypothesis scores maximum and the hypotheses with the maximum score are not equal, the perplexity of the hypotheses is calculated. To this end, we extended the particular text excerpt with 3 context words preceding and following the excerpt. For these text segments, we calculated perplexity with a pre-trained uncased German BERT model[11]. The hypothesis with the lower perplexity was selected.

The results of the experiments are presented in Table 4. Out of the two algorithms we applied on the data, better results could be achieved with the majority voting. The BLEU score improved by 2.9 points from 35.80 to 38.70 when compared to the result of the best model (`F-E2`).

[10] `F-E3` as a last-minute submission could not be used for ensembling

[11] https://github.com/dbmdz/berts#german-bert

Table 4: Ensembling results. The BLEU score achieved by each model separately and the BLEU score resulting from applying ensembling methods on the models' outputs (Majority Voting and Hybrid Ensembling)

F-E2	Jasper-PL	wav2vec XLSR-53 FT	MV	HE
35.80	32.97	30.39	**38.70**	37.62

7 Transcript Post-processing

Next to the Language Models for Speech Recognition, we evaluated an approach to using text-only data by training a supervised "spelling correction" (SC) model to correct the errors made by the STT model explicitly. Instead of predicting the likelihood of emitting a word based on the surrounding context, the SC model only needs to identify likely errors in the STT model output and propose alternatives. Intuitively, this task highly depends on the baseline model's quality: if the model transcribes very well, this task can be reduced to simply copying the input transcript directly to the output.

Most recent approaches for transcript post-processing use a transformer-based method: (Liao et al., 2021) use a modified RoBERTa structure and show an increase of 17.53 BLEU points on the self-augmented English Conversational Telephone Speech data set. On the LibriSpeech dataset, (Hrinchuk et al., 2019) show promising results using a pre-trained BERT as initialisation for their spell correction model, while (Guo et al., 2019) takes a different approach with a bidirectional LSTM.

We compared different Transformer architectures with their corresponding open-sourced pre-trained models and other post-processing methods.

The objective for all transformer models was set to next-sentence prediction (sequence to sequence generation) with a vocabulary size of 30'000, batch size of 16, and beam size for beam search set to 5. The models were initialised with pre-trained German embeddings and fine-tuned for up to 120'000 steps on the Shared Task training set described in 2.

- BERT (Devlin et al., 2018), having both encoder and decoder initialised with pre-trained weights.
- DistilBERT (Sanh et al., 2020), the lightweight alternative to BERT, reducing the training time up to 60%.

- ELECTRA (Clark et al., 2020), which uses a more sample-efficient pre-training approach for the encoder, called replaced token detection.

- SymSpell (Garbe, 2020),which is a spelling correction algorithm for correcting spelling errors based on Damerau-Levenshtein distances, stored in a pre-trained dictionary.

The following table shows the BLEU scores on the public test set, when performing post-processing on the output of the majority voting algorithm as described in 6. The Baseline refers to the BLEU score of the non-processed majority voting output.

Table 5: Post-processing BLEU scores on the public test set

System	Baseline	Post-processed
BERT	38.70	23.26
DistilBERT	38.70	26.66
ELECTRA	38.70	14.77
SymSpell	38.70	**30.65**

As the evaluations show, most post-processing attempts *decrease* the overall BLEU score, with SymSpell as the most straightforward approach performing best. Compared with previous work in this area, this could be explained by the limited amount of data available for training the transformer models. Due to lack of performance, we exclude the post-processing step in our final solution.

8 Conclusion

In this paper, we presented our contribution to the Shared Task "Swiss German Speech to Standard German Text" at SwissText 2021. Our solution combines the outputs of three models based on Fairseq, Jasper and Wav2vec XLSR-53 architectures. Because of time and resource constraints, we used only the labeled data set. Out of the 21 experiments we performed with the models, including transcript post-processing and ensembling, we achieved the best result by applying an ensembling method on the outputs of Fairseq model `F-E2` (BLEU score of 35.80) as the primary model, and `Jasper-PL` (32.97) and `wav2vec XLSR-53 FT` (30.39) as auxiliary models. We processed the three models' predictions with a majority voting algorithm and this way retrieved the most reliable candidate out of the provided translations for each utterance in the public test set. With this solution, we achieved a BLEU score of 39.39 on the private test set, which resulted in the third place out of four contributors in the competition.

Swiss German is a low-resource language, which makes training an STT or a Speech Translation system a challenging task. However, our experiments show that applying ensembling both on various models of the same architecture (as in Fairseq models `F-E1`, `F-E2` and `F-E3`) and on models based on various architectures (as implemented in our final solution) trained with limited data can lead to a score improvement of several BLEU points. Pseudo-labeling is another approach which contributes to model enhancement as we could observe with the `Jasper-PL` model. We will be further investigating these two methods aiming at improving the results despite the limited data currently available for Swiss German.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. Facebook AI.

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*.

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 *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. CEUR-WS.

Jiatao Gu Changhan Wang, Juan Pino. 2020. Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Eugen Dieth and Christian Schmid-Cadalbert. 1986. Schwyzertütschi dialäktschrift. *Sauerländer, Aarau*, 2.

Wolf Garbe. 2020. SymSpell: Fast spell correction algorithm.

Jinxi Guo, Tara N. Sainath, and Ron J. Weiss. 2019. A Spelling Correction Model for End-to-End Speech Recognition.

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. 2019. Correction of Automatic Speech Recognition with Transformer Sequence-to-sequence Model.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for End-to-End Speech Recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088. IEEE.

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. In *Proceedings of Interspeech 2019*, pages 71–75.

Junwei Liao, Yu Shi, Ming Gong, Linjun Shou, Sefik Eskimez, Liyang Lu, Hong Qu, and Michael Zeng. 2021. Generating Human Readable Transcript for Automatic Speech Recognition with Pre-trained Language Model.

H. Ney. 1999. Speech Translation: coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 1, pages 517–520 vol.1.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Thomas Hofmann Pelin Dogan-Schönberger, Julian Mäder. 2021. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2021. SwissText 2021 Task 3: Swiss German Speech to Standard German Text. In preparation.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Tanja Schultz, S. Jou, S. Vogel, and S. Saleem. 2004. Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems. In *INTERSPEECH*.

Beat Siebenhaar. 2003. Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats. *Linguistik online*, 15(3).

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

Matthias Sperber and Matthias Paulik. 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421.

John Richardson Taku Kudo. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.

Elvira Glaser Tanja Samardzic, Yves Scherrer. 2016. ArchiMob - A Corpus of Spoken Swiss German.

A. Waibel, A.N. Jain, A.E. McNair, H. Saito, A.G. Hauptmann, and J. Tebelskis. 1991. JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 793–796 vol.2.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. fairseq S2T: Fast Speech-to-Text Modeling with fairseq.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the Gap between Pre-Training and Fine-Tuning for End-to-End Speech Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9161–9168.

M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, and W. Ward. 1993. Recent Advances in Janus: A Speech Translation System. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, page 211–216, USA. Association for Computational Linguistics.

Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2020. Iterative Pseudo-Labeling for Speech Recognition. In *Proceedings of Interspeech 2020*, pages 1006–1010.