LEARNING TO TRANSLATE LOW-RESOURCED SWISS GERMAN DIALECTAL SPEECH INTO STANDARD GERMAN TEXT

Abbas Khosravani[1], *Philip N. Garner*[1], *Alexandros Lazaridis*[2]

[1]Idiap Research Institute, Switzerland
[2]Data, Analytics & AI Group — Swisscom AG, Switzerland

ABSTRACT

For a low-resourced language like Swiss German with no standard orthography and a significant variation in its written form, spoken language resources are more likely to come with translations than transcriptions. Moreover, the desired output of an automatic transcription system for Swiss German multi-dialectal speech is Standard German. This, in turn, is due to many applications that include our TV Box voice assistant and broadcast media. It follows that a translation is usually required as Swiss German and Standard German have mismatches on all linguistic levels. Unfortunately, there are not enough parallel text corpora available for training a proper translation system, nor enough in-domain speech translation (ST) data for training an ST system. We aim at investigating an end-to-end approach for multi-dialect Swiss German ST using transfer learning. Our ST model is based on an encoder-decoder architecture where we initialize the encoder with a cross-lingual speech representation model which is adapted to in-domain Swiss German speech data. We demonstrate that training the decoder on an out-of-domain ST corpus by preserving the encoder unit and then fine-tuning on in-domain ST data can be more effective than a cascade or vanilla direct ST.

Index Terms— Speech Translation, Swiss German, Multi-dialect, Transfer learning, Speech Recognition

1. INTRODUCTION

In German-speaking Switzerland people have the tendency towards informality in communication, which favors the use of Swiss German, a dialectal variety of the German language. Unlike many other countries, dialect is not a marker of low social class; everyone speaks Swiss German, but if they find cues that a person is not a dialect speaker they can switch to Standard German. Standard German is used in books, newspapers, and all official publications. Public radio and television have been more strictly in Standard German for many years, but local media are still in dialect.

Automatic speech recognition (ASR) of Swiss German is a considerable challenge owing to the lack of transcribed datasets and its considerable regional variation. There is no official orthographic convention for Swiss German varieties. The Dieth spelling system [1] (a phonetic transcription system) is usually used in scientific accounts for writing Swiss German dialects (referred to as GSW), but even phonetically identical words could be written differently. Moreover, due to many applications, including broadcast media and our TV box voice assistant, the desired output of an automatic transcription system is Standard German.

Swiss German is different from Standard German on all levels of linguistic analysis including vocabulary, pronunciation, orthography, and syntax. As a result, a machine translation (MT) system needs to be applied to the output of the ASR system to generate a Standard German translation of the utterance; it is termed a *cascade* speech translation (ST) system [2]. However, there is always a risk of error propagation and higher latency due to multiple inferences [3]. Unfortunately, Swiss German transcription is non-standard and hard to obtain, either for training ASR or MT systems. An *end-to-end* ST [4, 5] on the other hand, does not rely on the transcription of the source language for generating the translation. This as a result makes end-to-end ST systems more attractive.

Publicly available ST corpora for Swiss German speech to Standard German text are also limited. However, they are more likely to emerge due to the non-standardized written form of Swiss German dialects. Recently, an effort has been made to collect an ST corpus, called the Swiss Parliament Corpus, using a fully automatic procedure [6, 7]. They used a sentence-level forced alignment procedure using a Swiss German ASR model to align Standard German transcripts with Swiss German speech recording of the meetings in the parliament of Bern (mostly in Bernese dialect). Other examples of recent ST resources include the SRF Meteo weather report dataset [8] that consists of Swiss German weather reports of SRF Meteo with textual annotation in standard German, and the SwissDial corpus [9] with the speech in 8 different Swiss German dialects and both Swiss German and Standard German annotations.

We are interested in multi-dialect Swiss German speech translation into Standard German text for the Swisscom TV Box voice assistant. Due to the limited in-domain ST data, the use of separate ASR and MT corpora to train a cascade ST is currently unavoidable. In this paper, we investigate an

end-to-end approach for ST using transfer learning. Utilizing pre-trained components and multi-task learning are widely used methods to leverage ASR and MT datasets [5, 10]. Although pre-training is shown to be more effective than multi-task training due to sub-optimal solutions by the entire multi-task optimization problem [11], reusing pre-trained components also suffers from consistency issues [12].

We contribute by proposing a technique to train an end-to-end ST model which alleviates shortcomings in the existing methods. We exploit the powerful speech representation of self-supervised acoustic pre-training (wav2vec) [13] to address the low-resourced nature of the spoken dialects. Self-supervised pre-training is particularly efficient in low resource settings and it has been shown that fine-tuning pre-trained representations on the ST training data are beneficial [14]. Next, we leverage an out-of-domain ST dataset to pre-train the translation decoder in an encoder-decoder ST with attention. Direct optimization of the ASR output towards translation quality avoids semantic and length inconsistency compares to reusing a pre-trained MT decoder on text data [12]. Therefore, a more robust translation is expected to be achieved as we can pre-train the ST attention module, reduce the information loss and ASR error propagation. We empirically evaluate the proposed technique on our low-resourced in-domain ST dataset as well as on publicly available out-of-domain Swiss Parliament Corpus.

The rest of the paper is organized as follows. Section 2, introduces related work. In Section 3 we present our method for training an end-to-end ST by transfer-learning. We describe the data and baseline systems as well as experimental results in Section 4. Finally, we conclude in Section 5 with future directions.

2. RELATED WORK

With the recent success of end-to-end models for MT, ASR, and the availability of ST corpora, researchers start to explore end-to-end ST models. The cascade ST approach that divides the task into independent recognition and translation steps suffers from errors propagation from the ASR [15]. This in turn makes it challenging for machine translation which expects well-formed inputs. Following works turned increasingly to data-driven and statistical MT approaches to somewhat alleviated this issue and move from loosely-coupled cascade toward a tighter coupling [16, 2]. Researchers found an approximate integration over transcripts using *n*-best translation approach more tractable compared to a desirable full integration [17]. A follow-up work used word lattices as the interface between ASR and MT to improve translation performance when the weighted acoustic scores are incorporated into the MT unit [18]. In another approach, rather than trying to avoid early decisions using an *n*-best list or lattices, researchers introduced synthetic ASR errors to train a robust MT model [19, 12].

End-to-end approaches, on the other hand, learns a single model to map acoustic frames to target word sequence in the target language in a single step. Despite better performance obtained by the cascade approach [20], end-to-end methods are becoming more popular as they provide faster inference, rectifying error propagation (at least in theory), and minimizing the requirement for source language transcription. Recent investigation shows that in spite of data scarcity conditions which penalize the end-to-end approach, the two paradigms now perform substantially on par and the subtle differences observed in their behavior are not sufficient for humans neither to distinguish them nor to prefer one over the other [21].

Training an end-to-end ST model requires large-scale ST corpora which is difficult to obtain. Transfer-learning techniques including pre-training [22] and multi-task training [5] have been applied to leverage large-scale datasets available for ASR and MT. A recent finding indicates that utilizing pre-trained ASR and MT units and fine-tuning on weakly-supervised ST data generated using speech-to-text (TTS) synthesis models, strongly outperforms multi-task training [11].

An encoder-decoder ST model is usually pre-trained using an ASR encoder and an MT decoder following a fine-tuning step by multi-task learning which usually weights the losses of ASR, MT, and ST. It turns out that this approach suffers from consistency as both the encoder and decoder play different roles during pre-training and fine-tuning [12]. Moreover, a pre-trained MT on text data may weaken the model's sensitivity to prosody [3]. Recently, a Tandem Connectionist Encoding Network (TCEN) which consists of a speech encoder, a text encoder, and a target text decoder has been proposed to address this issue [12]. They pre-trained an ASR encoder model using Connectionist Temporal Classification (CTC) [23] loss function to generate embeddings in the source language and pre-trained an encoder-decoder MT model on large MT parallel corpora. Then they proposed solutions to solve the length inconsistency (ASR encoder output length is much larger than MT text input) and semantic inconsistency (ASR embedding space is different from MT word embedding space) of these two units when they are jointly trained by multi-task learning method. A knowledge distillation method has also been investigated recently in [24] to improve the end-to-end ST model by transferring the knowledge from a pre-train MT model. They first trained a text MT model and then an ST model is trained to learn its output probabilities through knowledge distillation.

3. METHOD

End-to-end speech translation aims to translate a sequence of speech features or raw waveform in source language defined as $\mathbf{x} = (x_1, x_2, ..., x_T)$ into a sequence $\mathbf{y^t} = (y_1^t, y_2^t, ..., y_{T''}^t)$ of units in target language vocabulary ($y_i \in \mathcal{V}^t$) without generating a sequence $\mathbf{y^s} = (y_1^s, y_2^s, ..., y_{T'}^s)$ of units in source language vocabulary ($y_i \in \mathcal{V}^s$). Unfortunately, directly train-

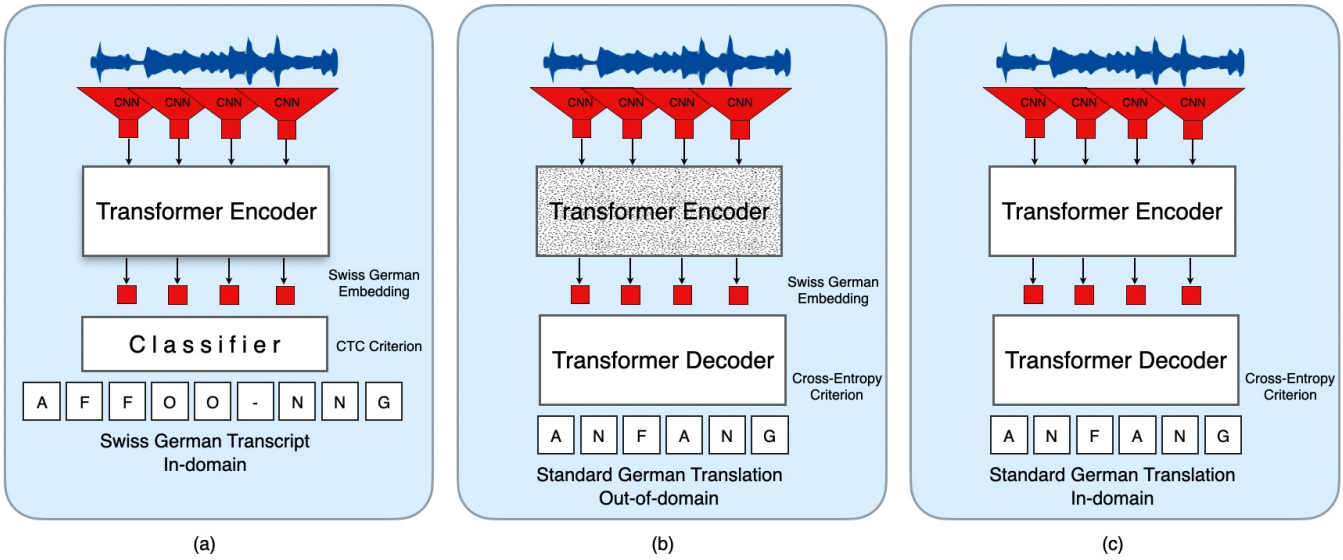

Fig. 1. The ST model architecture with training procedure. (a) ASR pre-training on in-domain speech data, (b) MT pre-training on out-of-domain ST data by freezing the ST encoder and (c) fine-tuning on in-domain ST dataset.

ing ST model is often impractical due to scarce end-to-end ST training corpora. This problem has been mainly tackled with data augmentation [11] and knowledge transfer [12] techniques. However, for a low-resource language like Swiss German with no standard orthography, there are not enough parallel text corpora available for training a proper MT model, nor enough in-domain ST data for a direct ST model. Our end-to-end ST model is based on an encode-decoder architecture with attention mechanism. We split the training procedure to pre-training and fine-tuning stages as illustrated in Figure 1. In pre-training stage, we initialize our ST encoder with a pre-trained ASR encoder as we describe in Section 3.1, but the ST decoder is pre-trained on an out-of-domain ST dataset $\mathcal{D}_{ST_{out}} = \{(\mathbf{x}_i, \mathbf{y}_i^t)_{i=1}^{P}\}$ which is described in Section 3.2. During fine-tuning, we use our low-resourced in-domain ST data $\mathcal{D}_{ST_{in}} = \{(\mathbf{x}_i, \mathbf{y}_i^t)_{i=1}^{P'}\}$ where $P' \ll P$, to train the ST model.

3.1. ASR pre-training

To train a Swiss German ASR model we use our in-domain multi-dialectal speech dataset $\mathcal{D}_{ASR} = \{(\mathbf{x}_i, \mathbf{y}_i^s)_{i=1}^{N}\}$ (see Section 4.1) [25]. The network architecture is based on the cross-lingual speech representations (XLSR) [26] model which is publicly available. It is built on top of wav2vec 2.0 [13], a framework for self-supervised learning of representations from the raw waveform of speech. Cross-lingual representation learning or pre-training aims to build models which leverage unsupervised multilingual data to share discrete acoustic units across languages, particularly, for low-resource languages, creating bridges across languages. It encodes speech audio via a multi-layer convolutional neural network which is then fed to a Transformer network [27] to learn contextualized representations via a contrastive task where the true latent is to be distinguished from distractors [13]. The XLSR model has a large capacity (315M parameters) and contains 24 Transformer blocks with model dimensions 1,024, inner dimension 4,096, and 16 attention heads. It is trained on 56k hours of speech data from 53 languages including the CommonVoice (36 languages, 3.6k hours) [28], Babel (17 languages, 1.7k hours) [29], and Multilingual LibriSpeech (8 languages, 50k hours) [30] corpora. We adapt the model by fine-tuning it on our multi-dialect speech data in the TV domain toward Connectionist Temporal Classification (CTC) objective [23]. To achieve this, we add a classifier on top of the model representing characters (including a word boundary token) in the source vocabulary $\mathcal{V}^s$ and train the model using the CTC loss function. Thus, there is no need for a decoder, and the ASR encoder will generate embeddings in the source language (after dropping the classification layer). For the first 10k updates only the output classifier is trained, after which the Transformer Network is also updated. Models are implemented in Fairseq [31].

3.2. MT pre-training

We initialize the encoder of the ST model with a the pre-trained ASR encoder as described in Section 3.1. Unlike similar studies that use a pre-trained MT model [12, 11], we pre-train our ST decoder using an out-of-domain ST dataset $\mathcal{D}_{ST_{out}}$ while preserving the ASR encoder (through freezing the weights during training). The ST decoder is based on 4 layers of Transformers blocks with model dimensions 512, inner dimension 2048, and 8 attention heads. The ST model is optimized using cross-entropy loss with label smoothing in

Fairseq.

Compared to using a pre-trained MT model on text data, this pre-training strategy avoids semantic and length inconsistency and optimizes the decoder towards translation. This in turn results in a more robust ST decoder, reduces the information loss and ASR error propagation as shown in Section 4.3. Moreover, the attention module is also pre-trained which is usually discarded during knowledge transfer (ASR and MT components model the source language differently). Note that, unlike many language pairs where large amounts of parallel text data are available, it is not easy to obtain such corpora for Swiss German text to Standard German text mainly due to non-standard orthography and significant variation in the written form.

3.3. ST fine-tuning

In the fine-tuning step, we use our in-domain ST dataset $\mathcal{D}_{ST_{in}}$ to adapt the ST model, which in turn strongly improves the translation quality (see Section 4.3). Unlike the previous step, the ST encoder is also adapted. The model is optimized using the same cross-entropy loss with label smoothing.

4. EXPERIMENTS

4.1. Data

Our in-domain ASR dataset $\mathcal{D}_{ASR}$ is a proprietary Swiss German multi-dialect dataset designed to improve Swisscom TV Box voice assistant [1]. It is designed to mimic a realistic usage scenario (e.g. users at different locations in a room talking to the TV box). With the help of 8 different microphones placed at different locations (2 close-talk, 1 pressure zone, 1 midfield, and 4 far-field) we collect user's voice commands when interacting with TV assistant [2], e.g. *'what will the weather be like tomorrow in Bern.'* in Swiss German. The recording sessions were either scripted (a speaker reads from a written note) or free talk (the speaker could improvise). The transcriptions are generated manually with the help of linguists in various Swiss German dialects. In total, the corpus consists of 392,420 short utterances comprising 440 hours of speech data from 3817 speakers (aged 14 to 89). It covers different Swiss German dialects from different regions including, Bern (BE), Valais (VS), Zurich (ZH), Eastern Swiss (EA), Grisons (GR), Central Swiss (CE), and Northwestern Swiss (NW), however, the distribution of the dialects is somehow imbalanced. The data is split into a train (417h) and validation (23h) sets for system development. We use a different set (23h) for system evaluation.

[1] https://www.swisscom.ch/en/residential/help/device/blue-tv/voice-assistant.html

[2] https://www.swisscom.ch/en/residential/plans-rates/inone-home/swisscom-blue-tv/functions.html

Our in-domain ST dataset $\mathcal{D}_{ST_{in}}$ is also a proprietary Swiss German multi-dialect dataset from Swisscom TV Box voice assistant but in the real scenario. In total, the corpus consists of 43,970 short utterances comprising 30 hours of speech data from 4147 speakers. Unlike $\mathcal{D}_{ASR}$, there is no dialect information, but annotations are available for each utterance in both Swiss German and Standard German. The data is split into a train (18h), validation (3h), and test (9h).

We use Swiss Parliament out-of-domain ST dataset $\mathcal{D}_{ST_out}$ which is an automatically aligned Swiss German speech to Standard German text corpus [7]. The dataset has been generated using a forced sentence alignment procedure to align the translation of the Swiss German speech recordings of meetings in the parliament of Bern (almost all in Bern dialect). The corpus includes an overall 293h of training data and 6h of testing data.

SwissDial is an annotated parallel corpus of spoken Swiss German dialects for 8 different regions of Aargau (AG), Bern (BE), Basel (BS), Graubunden (GR), Luzern (LU), St. Gallen (SG), Wallis (VS) and Zurich (ZH). The Standard German sentences from various topics are read and recorded by a limited number of speakers (one per dialect), and then manually translated into Swiss German dialects. It consists of 23195 utterances with parallel transcripts. We use this dataset for MT model training.

4.2. Baseline systems

4.2.1. Cascade ST

We build a cascade ST model as a baseline system by first training an ASR model for Swiss German as described in Section 3.1. Then we feed the predicted transcript from the ASR model as input to an MT model. The MT model is a simple character-based encoder-decoder network with Transformer layers which is trained on SwissDial [9] MT dataset $\mathcal{D}_{MT} = \{(\mathbf{y}_i^s, \mathbf{y}_i^t)_{i=1}^M\}$. The reason for using this model is that Swiss German and Standard German are closely related languages and the MT corpora are limited for this language pair. We split this dataset into train, validation, and test sets where the test set is a disjoint set of two Swiss German dialects.

4.2.2. Direct ST

A vanilla direct ST model with an encoder-decoder architecture is trained on both in-domain $\mathcal{D}_{ST_{in}}$ and out-of-domain $\mathcal{D}_{ST_{out}}$ datasets. The encoder is initialized with an XLSR model described in Section 3.1, but no fine-tuning has been performed to adapt it for Swiss German ASR. The ST decoder is the same Transformer network described in Section 3.2.

4.3. Results

We explore the effect of our pre-training strategy as described in Section 3 and compare it to the baseline systems. ASR quality is measured in terms of word error rate (WER) and character error rate (CER). For translation quality assessment we report the BLEU score [32] which is the scheme used by the annual Conference on Machine Translation (WMT)[3] with a tool denoted as SACREBLEU[4]. Its values (the higher the better) range from 0 to 100, but compare to the state-of-the-art systems in European language pair settings, the score for Swiss German text to Standard German text is usually much higher as they relatively closely related languages and the test set is in the same context as the development set.

We start by presenting our in-domain Swiss German multi-dialect ASR results on various dialects of $\mathcal{D}_{ASR}$ in Table 1 [25]. The results are obtained on a test set (23h) collected in a controlled environment with an average 17.4% WER and 4.18% CER. Lexicon-free beam-search decoding without any language model (LM) is used to generate transcriptions in Swiss German. VS (35h) and GR (44h) dialects have the lowest amount of speech data, so we observe higher WER, but for the eastern part of Switzerland (EA) we observe the lowest WER mainly due to more speech data that covers most of the variations. Owing to self-supervised pre-training using multilingual data, we achieve a high-performing multi-dialect ASR system in low-resource settings. In another experiment, we adapt the ASR model to our in-domain ST dataset by combining both training sets. Note that our in-domain ST dataset comes with both translation and transcription. This results in a huge drop in WER (CER) from 30.0% (10.6%) down to 12.5% (3.66%) on in-domain ST test set mainly due to acoustic mismatch between controlled and realistic scenarios and more simple or short utterances.

The MT model in our cascade ST baseline system achieved 45.9 BLEU and 31.1% WER on a disjoint set with two Swiss German dialects. Due to errors, inconsistencies, different context of MT corpus and unrecognized words associated with ASR output, we expected a drop in the performance of the MT model. As shown in Table 2, a BLEU score of 29.1 and WER of 36.2% were achieved for our baseline cascade ST model. The direct ST baseline model which is trained on the combination of in-domain (18h) and out-of-domain (293h) ST datasets could not outperform the cascaded approach. However, making a generalizable claim about the relative performance between the two ST models is not easy since we do not know how much increase in accuracy is achievable through the addition of more ST training data. The direct ST models exhibit an advantage over cascade ST which suffers from erroneous early decisions, however, it seems that lack of in-domain ST data and domain mismatch also play a key role.

[3]http://statmt.org

[4]https://github.com/mjpost/sacrebleu

Table 1. Performance of our multi-dialect ASR systems in terms of WER(%) on various Swiss German dialects. The results is reported on the test set of $\mathcal{D}_{ASR}$.

	GSW Dialects						
Evaluation	NW	BE	ZH	GR	EA	VS	CE
In-domain ASR	17.2	18.2	16.9	20.8	14.7	19.3	17.4

Table 2. Speech translation performance in terms of WER(%) and BLEU score on both in-domain and out-of-domain test sets using different systems.

	In-domain		**Out-of-domain**	
Systems	WER%	BLEU	WER%	BLEU
Baseline Cascade ST	36.2	29.1	–	–
Baseline Direct ST	45.3	21.1	55.8	16.1
Out-of-domain E2E ST	–	–	50.2	18.6
In-domain E2E ST	25.1	35.2	–	–

Following the training procedure described in Section 3 and illustrated in Figure 1, we train our end-to-end ST model. Table 2 shows the results on both out-of-domain and in-domain test sets before and after fine-tuning, respectively. The transfer-learned in-domain end-to-end ST model significantly outperforms both baseline systems. Freezing the pre-trained ST encoder would compel the model to generate intermediate embeddings in the source language. This in turn results in a pre-trained ST decoder when the model is trained on an out-of-domain ST dataset. Since the decoder is optimizing the ASR output toward the translation objective, the information loss and error propagation are minimized compared to initialization with a pre-trained MT decoder. Nevertheless, the result demonstrates that for low-resourced languages, especially those with considerable spoken variation, an end-to-end ST could be more effective than the cascade ST provided that we can have access to more out-of-domain ST datasets.

5. CONCLUSION & FUTURE WORK

We investigated an end-to-end approach for ST using transfer learning. The non-standard orthography of Swiss German and lack of parallel text corpora and for training an MT model as well as ASR error propagation limits the performance of the cascade ST. We leveraged an out-of-domain ST dataset to pre-train the ST decoder in an encoder-decoder ST architecture with attention. This approach which optimizes the ASR output toward translation avoids semantic and length inconsistency compared to reusing a pre-trained MT model.

Fine-tuning a pre-trained cross-lingual speech representation model was very effective for low-resource and highly dialectical Swiss German. We showed that our transfer-learned ST model can outperform both cascade and end-to-end ST baseline systems. For future work, we aim at investigating data augmentation and leveraging weakly supervised training data to improve end-to-end ST.

6. REFERENCES

[1] Eugen Dieth and Christian Schmid-Cadalbert, *Schwyzertütschi Dialäktschrift: Dieth-Schreibung*, vol. 1, Sauerländer, 1986.

[2] Hermann Ney, "Speech translation: Coupling of recognition and translation," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. IEEE, 1999, vol. 1, pp. 517–520.

[3] Matthias Sperber and Matthias Paulik, "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*, 2020, pp. 7409–7421.

[4] Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.

[5] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-sequence models can directly translate foreign speech," *Proc. Interspeech 2017*, pp. 2625–2629, 2017.

[6] Michel Plüss, Lukas Neukom, and Manfred Vogel, "Germeval 2020 task 4: Low-resource speech-to-text," in *SwissText/KONVENS*, 2020.

[7] Michel Plüss, Lukas Neukom, and Manfred Vogel, "Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus," *ArXiv*, vol. abs/2010.02810, 2020.

[8] Michael Stadtschnitzer and Christoph Schmidt, "Data-driven pronunciation modeling of swiss german dialectal speech for automatic speech recognition," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[9] Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann, "Swissdial: Parallel multidialectal corpus of spoken swiss german," *arXiv preprint arXiv:2103.11401*, 2021.

[10] Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6224–6228.

[11] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu, "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)*. IEEE, 2019, pp. 7180–7184.

[12] Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9161–9168.

[13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[14] Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Yannick Estève, and Laurent Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *Interspeech 2020*, 2020.

[15] Fred WM Stentiford and Martin G Steer, "Machine translation of speech," *British Telecom technology journal*, vol. 6, no. 2, pp. 116–122, 1988.

[16] Ye-Yi Wang and Alex Waibel, "Modeling with structures in statistical machine translation," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, 1998, pp. 1357–1363.

[17] Monika Woszczyna, N Coccaro, A Eisele, A Lavie, A McNair, T Polzin, Ivica Rogina, CP Rose, Tilo Sloboda, M Tomita, et al., "Recent advances in janus: A speech translation system," in *Third European Conference on Speech Communication and Technology*, 1993.

[18] Tanja Schultz, Szu-Chen Jou, Stephan Vogel, and Shirin Saleem, "Using word latice information for a tighter coupling in speech translation systems," in *Eighth International Conference on Spoken Language Processing*, 2004.

[19] Stephan Peitz, Simon Wiesler, Markus Nußbaum-Thom, and Hermann Ney, "Spoken language translation using automatically transcribed text in training," in *International Workshop on Spoken Language Translation (IWSLT) 2012*, 2012.

[20] Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, et al., "Findings of the iwslt 2020 evaluation campaign," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 1–34.

[21] Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi, "Cascade versus direct speech translation: Do the differences still make a difference?," *arXiv e-prints*, pp. arXiv–2106, 2021.

[22] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 58–68.

[23] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[24] Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong, "End-to-end speech translation with knowledge distillation," *Proc. Interspeech 2019*, pp. 1128–1132, 2019.

[25] Abbas Khosravani, Philip N. Garner, and Alexandros Lazaridis, "Modeling Dialectal Variation for Swiss German Automatic Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2896–2900.

[26] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[28] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[29] Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*. International Speech Communication Association (ISCA), 2014, pp. 16–23.

[30] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *Proc. Interspeech 2020*, pp. 2757–2761, 2020.

[31] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, June 2019, pp. 48–53, Association for Computational Linguistics.

[32] Matt Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels, Oct. 2018, pp. 186–191, Association for Computational Linguistics.