Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects

Clément Sicard, Kajetan Pyszkowski and Victor Gillioz

2 Models & Methods

During our whole experiment, we extensively used Hugging-Face to host our datasets, training results and models in different training stages. We also used multiple HuggingFace libraries to share datasets and train on multiple GPUs.

2.1 Datasets

We conducted a series of experiments for this paper with the four following datasets. Due to their respective particularities, they were used to different ends.

2.1.1 SwissDial

The SwissDial [1] dataset is an annotated parallel corpus of spoken Swiss German across eight major dialects (AG, BE, BS, GR, LU, SG, VS, ZH) with Swiss German and High German transcripts. It includes around 3 hours of high-quality audio per dialect. However, the dataset contains a class imbalance as it has around three times more Grisons Swiss German than any other dialect, which must be considered when training and evaluating models.

We randomly selected 20% of the data to constitute a test set. However, it must be noted that there is a single speaker per dialect, and most of the sentences are spoken across the eight dialects – for this reason, the independence of the train and test set is not perfect, and over-optimistic evaluation results are expected.

2.1.2 Swiss Parlament Corpus

The Swiss Parliament Corpus (SPC) [3] is a dataset of transcriptions of parliamentary speeches and proceedings from the Swiss National Council and Council of States. It consists of 293 hours of data. The corpus contains automatically aligned transcripts of speeches and proceedings from various parliamentary sessions, mainly in the dialect from Bern canton, with transcriptions in standard German. However, because of its size and the fact that there are almost only data points from Bern, in a juridic context, the usage of this dataset may introduce a bias in the training of a multi-dialect Swiss German ASR model. Moreover, the audio samples are often noisy, which brings additional difficulty for transcription – although Whisper is proved to be robust on noisy examples.

2.1.3 SDS-200

SDS-200 [2] dataset consists of 200 hours of speech data in Swiss German and corresponding transcriptions in standard German. The speech data was recorded from approximately 4k native speakers of Swiss German and covered various topics and Swiss German dialects. The speech data was recorded with a web tool open to the public. It covers a large part of the Swiss German dialect landscape. The dialect distribution roughly follows the speaker distribution in Switzerland.

The samples have been, for the most part, validated by the public. Moreover, the test set only contains speakers who are not present in the train set, and for which the speech audio has been sufficiently validated and is judged of high quality by the

Abstract

Recent breakthroughs in NLP largely increased the presence of ASR systems in our daily lives. However, for many low-resource languages, ASR models still need to be improved due in part to the difficulty of acquiring pertinent data. This project aims to help advance research in ASR models for Swiss German dialects, by providing insights about the performance of state-of-the-art ASR models on recently published Swiss German speech datasets. We propose a novel loss that takes into account the semantic distance between the predicted and the ground-truth labels. We outperform current state-of-the-art results by fine-tuning OpenAI's Whisper model on Swiss German datasets.
Keywords: ASR, Swiss German, sentence encoder

1 Introduction

Swiss German dialects are spoken by around 5 million people in Switzerland and are used in day-to-day life as a primary language of communication in the country. However, due to the non-existence of a standardized written form, standard German is used as the primary form of written communication outside of informal text messages. For this reason, we focus on transcription from multi-dialect Swiss German speech to standard German text, for which an extensive offer of language processing tools already exists.

This paper presents the results of evaluations we conducted on the recently available Swiss German datasets SwissDial [1], SDS-200 [2], and SPC [3]. Two models were used to conduct these evaluations: XLS-R [4], and Whisper [5] models. Both were evaluated and fine-tuned on Swiss German audio data with standard German annotations following transfer learning training strategies to respond to the low-resource situation of the Swiss German language.

Based on recent discussions[1] about the relevance of Word Error Rate (WER) [6], we additionally implement a semantic distance metric for evaluation by making use of language model embedding. We also use this metric to define a custom training loss to fine-tune Whisper model.

Our experiments show encouraging results for Whisper model. They suggest a soon-to-come availability of Swiss German ASR services, in a significant part due to the publication of annotated Swiss German speech datasets since 2021. We observe impressive results in Zero-Shot evaluations of Whisper models, and it even outperforms state-of-the-art results after training on a limited number of epochs. To the best of our knowledge, on the *multi-speaker-multi-dialect* SDS-200 dataset, we produce the current best results, with a WER of 20.6 and a BLEU of 66.6. We also achieve a better BLEU on the SPC dataset, namely 61.6. The model also offers high complexity in its outputs by including capital letters, numbers and punctuation with high accuracy.

---

[1] https://www.speechmatics.com/company/articles-and-news/the-future-of-word-error-rate

authors. We, therefore, consider it an essential set for model result evaluation.

2.1.4 Fleurs

Fleurs [7] is a speech dataset by Google which contains samples in 102 different languages, with approximately 12 hours of speech supervision per language. The dataset was built on top of the machine translation FLoRes-101 benchmark [8]. We use the standard German part of the dataset for evaluation.

2.1.5 ArchiMob

Another commonly referenced Swiss German audio dataset is ArchiMob [9]. It consists of 43 interview recordings in 14 different Swiss German dialects for around 70 hours of data. However, we decided against using this dataset in this project as it does not offer standard German transcription.

2.2 Models

In this project, we establish a baseline using Facebook's XLS-R 1B [4] and further train OpenAI's Whisper [5] as our primary focus. As Whisper's medium version has 769M parameters and its large version has 1.5B, we believe that the comparison between the two architectures is reasonable. Furthermore, to compute our custom loss $\mathcal{L}_{\mathsf{SemantiX}}$, which requires the use of a LLM, we use a pre-trained XLM-RoBERTa [11] for multilingual sentence embeddings [12]. However, as this model is not the main focus of this research, we do not present it in detail here.

2.2.1 Facebook's XLS-R 1B

XLS-R [4] is a multilingual speech representation learning model. It is based on Wav2Vec2 [13] self-supervised learning framework and was pre-trained on 436k of unlabelled speech from 128 languages. No Swiss German data was used for training. The most recent publications on Swiss German ASR [2] [10] showed XLS-R capacity for Swiss German speech to standard German transcription. Following these publications, we similarly focused on the 1B parameters model (XLS-R 1B).

We implemented this model to offer a basis for comparison to Whisper results when trained in a similar setting. Moreover, results on the SwissDial dataset have yet to be published. For this reason, the training and evaluation of the XLS-R 1B model allowed us to set a first benchmark and provide a better perspective on our results.

To respond to the low-resource setting of Swiss German and the limited scope of this research, we use a model that has already been fine-tuned on multiple German speech datasets [14]. It also offers a CTC [15] beam search decoder language model, which improved our overall results. The output vocabulary consists of lowercase letters and a limited number of special characters. This model reached a WER [6] of 8.13, and a CER [16] of 2.18 on the Common Voice 8 [17] test set when combined with its language model decoder.

2.2.2 OpenAI's Whisper

OpenAI's Whisper [5] architecture is a collection of multiple sizes of a single model designed to bridge the gap between

small supervised models trained on limited data and large unsupervised models that require precise fine-tuning to perform specific tasks. For example, Whisper models were trained on speech recognition tasks, such as speech-to-text and language identification. Using a large amount of weakly-labelled data, the models can learn from various speech patterns and variations, resulting in improved robustness and generalization.

The training process for Whisper is scaled to 680k hours of multilingual and multitask supervision, resulting in models that can generalize well to standard benchmarks and are often competitive with fully supervised models without the need for fine-tuning. For example, on the Common Voice [17] benchmark, Whisper models achieved a word error rate of 4.5%, which significantly improved over the previous state-of-the-art.

Of the 680k hours of audio used for training, 117k hours cover 96 different languages, and the dataset also includes 125k hours of language-to-English translation data. The model's capabilities and results on multiple datasets suggest that Whisper could be the next state-of-the-art in Swiss German ASR tasks. With access to a growing number of high-quality Swiss-German datasets, there is potential to create a new baseline for future Swiss German ASR tasks.

Additionally, as OpenAI's Whisper architecture includes model sizes ranging from tiny (39M parameters) to large (1.5B parameters), exploring how the different model sizes perform on the same tasks is an exciting way to explore.

2.3 Inviting Semantics to Traditional ASR Metrics and Losses

To evaluate our results, we use Word Error Rate (WER) [6] and Character Error Rate (CER) [16] metrics. Similarly to recent Swiss German ASR literature, we also compute the BLEU [18] metric, which is standard in the automatic evaluation of machine translation. Indeed, our task is, to some extent, a translation task due to the particularities of Swiss German with respect to standard German, such as differences in verb conjugation, vocabulary or morphology. Examples of these differences can be found in the result section.

2.3.1 Motivation to Involve Semantics

Error-rate metrics are predominant metrics in the world of ASR. However, as described by Kim et al.[19], only using WER can be misleading, as it only takes into account literal correctness instead of semantic correctness. As shown in the paper, their proposed metric SemDist takes into account the semantics of the predicted sentence by leveraging the power of LLMs. It also inspired the creation of a new custom loss, $\mathcal{L}_{\mathsf{SemantiX}}$, which we will describe in 2.3.3. Let us first introduce semantic distance as a metric and then explain how we used it to derive a new loss.

2.3.2 Semantic Distance as a Metric

Semantic distance uses the fact that LLMs can create word/sentence embeddings, which can be seen as point encoding in a latent semantic space. One can then compute the "distance" between those points using a score like cosine similarity. Facebook AI's paper by Kim et al. [19] does precisely that by introducing SemDist. It is defined as follows, with **x**

Models	SwissDial			SDS-200			SPC			Fleurs		
	WER	CER	BLEU	WER	CER	BLEU	WER	CER	BLEU	WER	CER	BLEU
XLS-R 1B	69.3	33.8	12.3	76.6	40.4	8.8	73.6	39.9	12.7	10.2	5.6	71.3
Whisper Tiny	80.2	42.3	10.3	92.7	54.9	6.61	96.3	55.7	9.7	37.5	12.9	46.0
Whisper Base	66.8	33.2	19.0	78.0	43.4	13.8	70.4	37.8	20.3	25.6	8.4	60.2
Whisper Small	46.8	22.6	36.4	51.0	27.3	34.2	50.9	27.1	37.3	14.7	4.1	75.0
Whisper Medium	33.8	16.9	50.8	36.8	20.4	49.7	37.4	20.2	50.8	10.5	2.9	81.7
Whisper Large	**29.4**	**14.8**	**56.2**	**31.7**	**18.0**	**55.6**	**33.2**	**18.2**	**55.6**	**8.7**	**2.3**	**84.7**

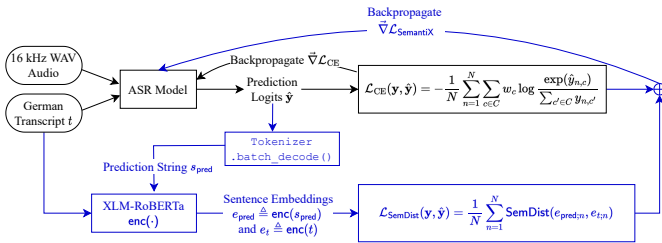Figure 1: Baseline performances on a Zero-Shot evaluation

Figure 3: Flowchart of training pipeline – the blue elements of the flowchart are needed for $\mathcal{L}_{\mathsf{SemantiX}}$. The black elements are needed for training with the initial cross-entropy loss.

and $\mathbf{y}$ two sentence embeddings encoded by a sentence encoder [20].

$$\mathsf{SemDist}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^\top \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||} \qquad (1)$$

A result closer to zero indicates a higher semantic similarity and vice versa. We compute $\mathsf{SemDist}(\cdot, \cdot)$ during evaluation to add a level of comprehension to our results.

From this sample metric, we also derive a loss $\mathcal{L}_{\mathsf{SemDist}}$ on a batch of size N, with $\mathbf{y} = \{y_n\}_{n=1}^N$ and $\hat{\mathbf{y}} = \{\hat{y}_n\}_{n=1}^N$ respectively batches of labels and predictions as the mean of the respective distances :

$$\mathcal{L}_{\mathsf{SemDist}}(\mathbf{y}, \hat{\mathbf{y}}) \triangleq \frac{1}{N} \sum_{n=1}^N \mathsf{SemDist}\{\mathsf{enc}(y_n), \mathsf{enc}(\hat{y}_n)\} \qquad (2)$$

In our setup, $\mathsf{enc}(\cdot)$ corresponds to the sentence encoded by XLM-RoBERTa [11] [20].

2.3.3 SemantiX Loss

By default, Whisper uses a cross-entropy loss between the undecoded outputs of the model and the tokenized ground truth labels to quantify error. It is defined as follows, with C the set of all possible tokens that the model can output:

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{n=1}^N \sum_{c \in C} w_c \log \frac{\exp(\hat{y}_{n,c})}{\sum_{c' \in C} y_{n,c'}} \qquad (3)$$

We introduce a new loss $\mathcal{L}_{\mathsf{SemantiX}}$, which uses both the well-known cross-entropy loss for literal correctness but also $\mathcal{L}_{\mathsf{SemDist}}$ for semantic correctness. The loss takes two hyperparameters $\alpha, \beta \in \mathbb{R}_+$. Combined, it gives the following function:

$$\mathcal{L}_{\mathsf{SemantiX}}(\mathbf{y}, \hat{\mathbf{y}}) \triangleq \alpha * \mathcal{L}_{\mathsf{SemDist}}(\mathbf{y}, \hat{\mathbf{y}}) + \beta * \mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \hat{\mathbf{y}}) \qquad (4)$$

It enables us to weigh the importance of the meaning of the prediction in addition to outputting the right tokens. In our experiments, we also used a variant of it with a product instead of a sum:

$$\mathcal{L}'_{\mathsf{SemantiX}}(\mathbf{y}, \hat{\mathbf{y}}) \triangleq (\gamma + \mathcal{L}_{\mathsf{SemDist}}(\mathbf{y}, \hat{\mathbf{y}})) * \mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \hat{\mathbf{y}}) \qquad (5)$$

Both implementations use `CosineEmbeddingLoss` from PyTorch [21].

2.4 Training Setup

2.4.1 XLS-R

A Zero-Shot evaluation was first performed with the German pretrained XLS-R 1B model on the following datasets: SwissDial, SDS-200, SPC and Fleurs. In a second time, we fine-tuned the model on the SwissDial and SDS-200 datasets, following the procedure suggested by the authors[2]. We used the described hyper-parameters with a learning rate of $1e - 4$ and trained for 5 epochs.

2.4.2 Whisper

Similarly to XLS-R 1B, we evaluated Whisper using a Zero-Shot approach on the following datasets: SwissDial, SDS-200, SPC and Fleurs. The first three datasets were used to baseline Whisper's initial performance on Swiss German. The last dataset was used as a control dataset to see how well our Whisper implementation's results aligned with Whisper's paper's results. Moreover, we evaluated all the different sizes of Whisper to compare them against each other.

Furthermore, we mostly fine-tuned the medium size of Whisper as it offered the best performance/training-time ratio. In addition, we fine-tuned Whisper on multiple combinations of three Swiss German datasets (see 2). As Whisper training time is consequent and we only fine-tuned it, we kept the number of training epochs close to one for each dataset. The different outcomes and conclusions are described in 3.1. Furthermore, we also fine-tuned Whisper using our innovative loss, described in 2.2.2 (see Fig. 2). Our fine-tuning method was heavily inspired by the procedure described by Sanchit Ghandi[3].

3 Results

3.1 Baselines

Our first experiment was a Zero-shot evaluation on the Swiss-Dial test set using the dialectal transcripts. As shown in Figure 5, the results are quite bad. This can be explained by the lack of a standardized written form of the Swiss German language and the fact that Whisper has not been introduced to any sort of Swiss German transcript, which renders it close to impossible to produce a correct output. These results and a secondary evaluation with standard German transcripts reported in Figure 1 confirmed our decision to use those for the rest of the experiments.

In a second time, we evaluated XLS-R and every size of Whisper, in a similar Zero-Shot [22] fashion, on each of our datasets with standard German transcripts. We include the Fleurs dataset to assess how well each of the model performs on standard German. All these results are reported in Figure 1. Even though Whisper Large performs slightly better on SDS-200 and Fleurs, we decided to focus on Whisper Medium model for further fine-tuning and for computational reasons. A huge improvement is still noticeable between different sizes of Whisper, and XLS-R seems to perform worse on Zero-Shot evaluation. The well known ability of Whisper to generalize to unknown data and regional accents might explain these results.

---

[2] `https://huggingface.co/blog/wav2vec2-with-ngram`
[3] `https://huggingface.co/blog/fine-tune-whisper`

Models	SwissDial			SDS-200			SPC			Fleurs		
	WER	CER	BLEU	WER	CER	BLEU	WER	CER	BLEU	WER	CER	BLEU
Plüss et al. [2] (norm.)	-	-	-	21.6	-	64.0	-	-	-	-	-	-
Plüss et al. [10] (norm.)	-	-	-	-	-	-	23.7	-	60.7	-	-	-
XLS-R 1B	17.7	10.3	62.9	25.2	16.4	53.6	40.2	24.0	37.9	47.9	25.8	25.0
Whisper ($\mathcal{L}_{\mathrm{CE}}$)	**14.3**	**7.5**	**77.7**	21.2	**12.9**	65.4	28.9	**16.0**	**61.6**	**16.6**	**5.4**	72.9
Whisper ($\mathcal{L}_{\mathsf{SemantiX}}$)	21.2	9.8	65.3	**20.6**	13.0	**66.6**	30.4	18.6	56.7	**16.6**	5.5	**73.0**

Figure 2: Performance of our models on a post-training evaluation

Ground Truth	Prediction	WER	CER	BLEU	SemDist
Andererseits seien strategische Entscheide für den Rückgang verantwortlich.	Andererseits sei ein strategischer Entscheid für den Rückgang verantwortlich.	50.0	5.3	41.1	1.0
Boeing lehnte eine Stellungnahme ab.	Boeing hat den Stellungnahme abgelehnt.	60.0	38.9	0.0	1.1
Wegen des Brandes war die Dorf**strasse** für mehrere Stunden gesperrt.	Wegen des Brandes war die Dorf**strafe** mehrere Stunden gesperrt.	20.0	10.0	59.5	23.3
Aus Syrien stammten im Mai 52 Asyl**bewerber**.	Aus Syrien stammten im Mai 52 Asyl**werber**.	14.3	4.7	70.7	0.2
Inzwischen ist es kurz vor 22 Uhr.	Mittlerweile ist es kurz vor 10 Uhr.	28.6	38.2	41.1	23.4

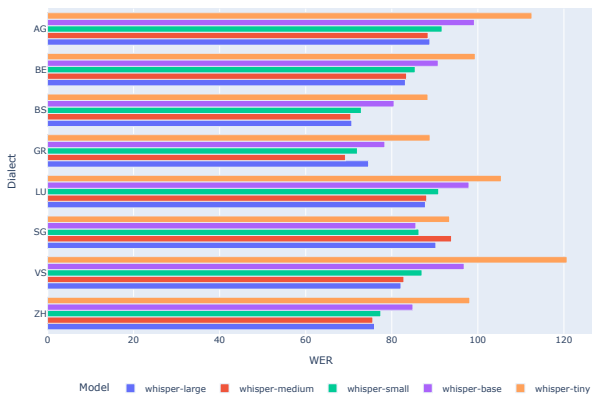Figure 4: Selected examples of predictions. It highlights the quality of SemDist as a metric

Figure 5: Zero-Shot evaluation results of different Whisper sizes on the SwissDial dataset, with dialectal transcripts,

3.2 Our Models

Based on the baseline results, we decided to fine-tune Whisper Medium on our datasets, leveraging the semantic losses (Fig. 2). Whisper ($\mathcal{L}_\mathsf{SemantiX}$) has been trained first on a shuffled concatenation of SwissDial and 20% of SPC (to have roughly as many samples from both) using cross-entropy loss for 1 epoch over the whole train dataset ($\approx$ 60k samples), then trained for another round on SDS-200 for 2 epochs on the full train dataset ($\approx$ 37k samples), this time using $\mathcal{L}_\mathsf{SemantiX}$. Whisper ($\mathcal{L}_\mathrm{CE}$) has been trained in the same setting but sticking to $\mathcal{L}_\mathrm{CE}$ the whole training. Both models were trained with a batch size of 10 and 2 gradient accumulation steps – depending on the runs, on NVIDIA A100 and RTX Quadro 6000 GPUs. In this setup, we obtain better results than state-of-the-art on SDS-200 with Whisper ($\mathcal{L}_\mathrm{CE}$) for CER and with Whisper ($\mathcal{L}_\mathsf{SemantiX}$) for BLEU and WER.

We observe that after training Whisper on any of our datasets, it tends to suffer from catastrophic forgetting, as we see, for instance, in the performance decrease for the Fleurs dataset, between fine-tuned models results and the Zero-Shot ones. To avoid this, the models should probably have been trained on all the datasets altogether for 2 to 3 epochs in a similar fashion as in the Whisper paper [5].

Note that Whisper – as opposed to XLS-R 1B – works with unnormalized text – namely, it preserves punctuation, casing, digits, onomatopoeia, etc – for both input and output. To our understanding, evaluation metrics in Swiss German ASR literature, namely WER and BLEU, were computed in a normalized setting. The transcription contained the characters `a-z`, ä, ö, ü, and spaces, with no punctuation, casing, and numbers spelt out. We computed our metrics on normalized predictions and ground truth but observed only minor differences, suggesting that Whisper interpolates punctuation and casing very well. In some examples presented in Figure 4, we observe that Whisper encounters difficulties in transposing verb tenses from the Swiss German conjugation to Standard German. Probably due to such differences – WER results might be too pessimistic. The 2nd sentence is a good example, where we observe a high WER but a low SemDist, because the meaning

of the predicted sentence remains very close. Therefore, we find that even though WER might seem discouraging, the overall $\mathcal{L}_\mathsf{SemDist}$ is very low on our models for all of our datasets (Fig. 6) and suggests good transcription capacities by our model. We performed a *p*-value test between the standard metrics and SemDist on our model predictions and observed no significant correlation. Hence, SemDist seems to be an interesting new metric that could bring a new level of understanding in ASR transcription tasks.

We believe that semantic distance could significantly improve ASR task evaluation when combined with error-rate measures. Additional examples can be found in Figure 4.

	SwissDial	SDS-200	SPC	Fleurs
Whisper Medium	9.2%	11.0%	8.8%	2.5%
Whisper ($\mathcal{L}_\mathrm{CE}$)	5.0%	4.9%	8.1%	5.7%
Whisper ($\mathcal{L}_\mathsf{SemantiX}$)	4.4%	4.2%	6.7%	5.3%

Figure 6: Average semantic distance $\mathcal{L}_\mathsf{SemDist}$, as a metric, during evaluation on our fine-tuned models

4 Discussion

Using the large version of Whisper for Zero-Shot evaluation on our datasets also showed promising results. Therefore, one should try our experiments with the large version of Whisper if one has the computing capacities.

Nevertheless, an observed downside of the Whisper fine-tuning is the catastrophic forgetting of our model. Indeed, after training, one can see that the models perform worse even in German. We believe however that, with enough computing capacities, this catastrophic forgetting could be attenuated while conserving the encouraging training results presented in this paper.

Another important finding of this study is that traditional metrics might not be sufficiently extensive to capture all the intrinsics of the "Swiss German audio to High German text" task. The shortcomings of traditional metrics and losses opened the possibility to have predicted sentences to be further processed, using for instance a LLM to match the conjugation rules.

Summary

This study shows that the newly released Whisper model is able to impressively generalize its knowledge to unseen languages such as Swiss German dialects, despite important disparities between speech and transcription. Moreover, fine-tuning Whisper on a diverse set of Swiss German datasets can significantly improve its overall performance. As presented before, we were able to fine-tune Whisper and outperform state-of-the-art using SwissDial, SDS-200, and to a limited extent SPC. These results were obtained with minimal hyperparameters tuning and a small number of epochs. As a first step in this direction, we created our custom loss function $\mathcal{L}_\mathsf{SemantiX}$, a mix between traditional and semantic metrics. However, we did not get consistently better results in traditional metrics such as WER, BLEU and CER. Nevertheless, our custom loss function helped reduce the semantic distance between our predictions and transcripts on Swiss German datasets.

References

[1] Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German. *CoRR*, abs/2103.11401, 2021.

[2] Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. SDS-200: A Swiss German speech to standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France, June 2022. European Language Resources Association.

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

[4] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale, 2021.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. Technical report, OpenAI, 2022.

[6] Ye-Yi Wang, A. Acero, and C. Chelba. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 577–582, 2003.

[7] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech, 2022.

[8] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021.

[9] Tanja Samardžić, Yves Scherrer, and Elvira Glaser. Archi-Mob - 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, May 2016. European Language Resources Association (ELRA).

[10] Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. Swiss german speech to text system evaluation, 2022.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2019.

[12] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[13] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[14] Jonatas Grosman. Fine-tuned XLS-R 1B model for speech recognition in German. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-german, 2022.

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

[16] Andrew Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. 01 2004.

[17] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[19] Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. Semantic distance: A new metric for asr performance analysis towards spoken language understanding, 2021.

[20] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[22] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. *CoRR*, abs/1703.05175, 2017.