Interspeech 2022
18-22 September 2022, Incheon, Korea

Comparison of Unsupervised Learning and Supervised Learning with Noisy Labels for Low-Resource Speech Recognition

Yanick Schraner, Christian Scheller, Michel Plüss, Lukas Neukom, Manfred Vogel

University of Applied Sciences and Arts Northwestern Switzerland

yanick.schraner@fhnw.ch, christian.scheller@fhnw.ch, michel.pluess@fhnw.ch, lukas.neukom@fhnw.ch, manfred.vogel@fhnw.ch

Abstract

Supervised training of end-to-end speech recognition systems usually requires large amounts of transcribed speech data to achieve reasonable performance. This hinders its application to problems where the availability of annotated data is low since manual labeling is costly. Often, however, large amounts of speech data with imperfect transcriptions are available, which can be automatically aligned to generate noisy labels. In this work, we compare how supervised learning on noisy data from forced alignment compares to semi-supervised learning and self-supervised representation learning. The latter two have shown great success in improving speech recognition using unlabeled data. We employ noisy student training for semi-supervised learning and wav2vec 2.0 for self-supervised representation learning. We compare these methods on 2324 hours of Swiss German audio with automatically aligned Standard German text. Using speech data with noisy labels for supervised learning leads to a word error rate (WER) of 26.4% on our test set. Using the same data for wav2vec pretraining leads to a WER of 27.8%. With noisy student training, we achieve a WER of 30.3%.

Index Terms: speech recognition, low-resource, speech translation, semi-supervised, self-supervised, forced alignment

1. Introduction

Supervised training for automatic speech recognition (ASR) has shown remarkable success in recent years [1, 2]. However, these state-of-the-art models require a large corpus of labeled speech data to work well. Acquiring clean per-sentence transcriptions of speech data is expensive. This hinders the application of these models to domains where labeled data is scarce.

Semi-supervised and self-supervised methods utilizing large amounts of unlabeled data have recently shown great success in mitigating this problem by drastically reducing the required amount of labeled data [3, 4, 5]. Another approach is the forced alignment of publicly available transcribed speech recordings to create a labeled ASR corpus [6, 7]. However, such transcriptions are often imperfect and contain added, missing, or altered parts, which leads to noisy labels.

In this work, we compare the effectiveness of state-of-the-art semi-supervised and self-supervised ASR methods to supervised training on noisy labels obtained from forced alignment. For semi-supervised training, we employ noisy student training for automatic speech recognition [8]. The noisy student method was first introduced by Xie et al. [9] for on computer vision tasks. For self-supervised training, we use wav2vec 2.0 [3]. Both methods demonstrate state-of-the-art performance on benchmarks such as LibriSpeech, Common Voice (various languages) and TIMIT [3, 4, 10]. To create the training data for supervised training, we employ the forced alignment procedure proposed by Pluess et al. [7] to get sentence-level audio-text-pairs.

We conduct this research on the problem of Swiss German ASR that is concerned with the transcription of Swiss German Speech to Standard German text. Swiss German ASR can be viewed as a speech translation (ST) task with similar source and target languages [7]. This problem poses an ideal testbed for our comparison since the amount of publicly available labeled data is low. To our knowledge, there are three public datasets with Swiss German audio and Standard German transcripts available. The Radio Rottu Oberwallis dataset [11] and the SwissDial dataset [12] combined consist of 28 hours of audio with clean labels. The Swiss Parliaments Corpus (hereafter SPC-public) [7] consists of 293 hours of audio with noisy labels. In contrast, speech recordings with approximate transcripts are available in large amounts (e.g. recorded parliament sessions [13]). In addition to the word error rate (WER), we report the BLEU [14] score as this is common in ST tasks.

Existing ASR approaches for Swiss German predominantly use supervised methods [11, 15, 7, 13]. Due to the low-resource nature of Swiss German and the scarcity of labeled training data, we argue that the use of additional unlabeled data could lead to considerable improvements on the task at hand.

The main contribution of this research is a comparison between supervised training on noisy labels, noisy student training, and wav2vec 2.0 on the example of Swiss German ASR. To our knowledge, we are the first to carry out such a comparison. We argue that our results are language-independent and can be transferred to other datasets with similar characteristics. To this end, we propose an evaluation protocol where approximately transcribed speech recordings are used either as unlabeled data for semi-supervised and self-supervised methods or, after a forced alignment step, as noisily labeled data for supervised training.

The remainder of this paper is structured as follows: section 2 describes data sources and processing, section 3 provides further detail on the adopted methods, and section 4 outlines the experiments carried out.

2. Data

As outlined in the introduction, we conduct this research on the problem of transcribing Swiss German speech to Standard German text. We consider this problem to be low-resource since the amount of publicly available, cleanly labeled data is comparatively low, as highlighted in Table 1.

To compare supervised learning with noisy labels to noisy student training and wav2vec 2.0, we require a corpus of speech recordings paired with an approximate Standard German transcription. The approximate transcriptions are necessary to generate noisy labels for supervised learning. To this end, we ob-

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10620

tained a corpus of audio recordings and minutes of parliamentary sessions from the cantons of Appenzell Ausserrhoden and Obwalden and the city of Bern. Due to data privacy concerns, this data is not public. Neither audio recordings nor minutes are perfect and therefore provide an ideal data source for this work. The audio recordings are not of studio quality. They contain background noises and minor mistakes in speech. The minutes can exhibit errors such as missing, added, and altered parts. This private corpus, which we subsequently refer to as SPC-private, contains 1864 hours of raw audio recordings in total.

As labeled data for noisy student training and wav2vec 2.0, we use the Swiss Parliaments Corpus [7]. This corpus provides 293 hours of noisily labeled training data and a cleaned test set. The labeling process for this corpus was the same as adopted in this work, outlined in Section 3.1.

For language model training, we combined europarl [20], news-crawl 2019 [21], parlspeech v2 [22], and SPC-public and SPC-private train texts to obtain a total of 68 Million German sentences.

3. Methods

3.1. Supervised Learning with Noisy Labels

We follow the forced alignment procedure proposed by Plüss et al. [7] to create sentence-level alignments of the raw audio data from the parliament speeches and the corresponding minutes. In this procedure, we first transcribe the audio using a Standard German ASR model to obtain low-quality transcripts including per-word timestamps of the corresponding audio segment. The ASR transcripts are then aligned with the minutes using the Needleman-Wunsch algorithm. We then create an alignment of each sentence from the minutes to the audio using the ASR timestamps and obtain a sentence-level training corpus. Finally, we train Transformer networks in a supervised learning setup on this data.

3.2. Noisy Student Training

Noisy student training is a semi-supervised learning approach. Park et al. [8] have applied the noisy student training scheme successfully to train models for ASR. We train a teacher model on labeled speech data in the first step. We then use the teacher to generate pseudo labels on unlabeled speech data. In a third step, we train a student model on the combination of labeled and pseudo labeled speech data. This process can be repeated several times, with the student becoming the new teacher, generating more accurate pseudo labels.

Lang	Corpora	Size
EN*	MLS† [16], CV [17], LS† [6], SWC [18]	48′201h
DE*	MLS† [16], CV [17], SWC [18]	3′415h
SG	SD [19], RRO [11]	28h

Table 1: *Publicly available cleanly labeled ASR datasets for English (EN), German (DE), and Swiss German (SG). *For English and German, we show only a summary of the largest available corpora. †We consider MLS and LS as cleanly labeled since the transcripts used for the forced alignment are exact.*

Dataset	Dialect	Raw	Annotated
SPC-public	BE	460h	293h
SPC-private	AR, BE, OW	1864h	1066h
Total	AR, BE, OW	2324h	1359h

Table 2: *Summary of the training data used in this work.*

3.3. Wav2vec 2.0

For self-supervised learning, we employ the wav2vec 2.0 framework [3] that learns general representations from unlabeled speech data followed by fine-tuning on transcribed data. These models consist of a convolutional feature encoder that maps audio features into latent space, followed by a Transformer encoder that produces representations. During finetuning, a classification head maps representations to target vocabulary tokens.

4. Experiments

4.1. Data

The data sources used in our experiments are summarized in Table 2. Through the labeling process described in section 3.1, we obtained a total of 1066 hours of noisily labeled speech to text data for the SPC-private corpus. We provide examples of the outputs from the labeling process in Table 3. With SPC-public and SPC-private, we obtained 2324 hours of unannotated and 1359h of annotated speech-to-text data. We describe how we used this data for each experiment in the following.

4.2. Baseline

As a fair baseline, we train supervised Transformer ASR models on the SPC-public dataset without any additional data.
Data. The baseline is trained on the SPC-public dataset. We do not apply a filter based on the Intersection over Union (IoU) of the forced alignments. A speed perturbation [23] of 0.9 and 1.1 is applied to all audio files.
Training. We use the ESPnet2 toolkit [24] to train our baseline model. The model is based on the transformer configuration provided in the Common Voice ASR recipe [1]. We trained for a maximum of 100 epochs or stopped training when the last three epochs did not improve the validation accuracy.
Language Model. We train an LSTM based language model with two layers and a hidden size of 650 units. We trained for a maximum of 20 epochs on the language model data described in section 2. The training is stopped when the last three epochs did not improve the validation accuracy. This language model is also used for supervised learning with noisy labels and noisy student experiments.

4.3. Supervised Learning with Noisy Labels

The supervised learning experiments rely on the baseline and use the same network architecture, training hyperparameters, language model and decoding parameters.
Data. The supervised learning model is trained on the SPC-public and SPC-private datasets with noisy labels. We apply a speed perturbation of 0.9 and 1.1 to all audio files and do not filter out samples based on the IoU of the forced alignments.
Training. We train the supervised learning model exactly the same way as the baseline is trained.

[1] https://github.com/espnet/espnet/tree/master/egs2/commonvoice/asr1

Method	Sentence
Manual	Das gilt ganz besonders für heikle Themenbereiche wie dem Sozialrevisorat und -inspektorat, auch wenn in der Unterstellung nicht alles ganz genau so herauskommt wie es der Ausschuss das wollte. Allerdings
Aligned	Ganz besonders gilt das für heikle Bereiche wie dem Sozialrevisorat und dem Sozialinspektorat, auch wenn nicht alles ganz genau so herauskommt wie es der Ausschuss empfohlen hat.
	Editdistance: 13 (43.3% WER)
Manual	Wir sind der Meinung, dass eigentlich das Volk auch hier sagen sollte, wie viel Geld zur Verfügung gestellt wird.
Aligned	Wir sind der Ansicht, dass auch hier das Volk bestimmen sollte, wie viel Geld zur verfügung gestellt wird.
	Editdistance: 7 (36.8% WER)
Manual	Gebt euch künftig doch bitte wieder etwas Mühe, entschuldigung, danke. Dann hat als nächstes Franziska Grossenbacher als Einzelvotantin das Wort. Ja, danke Tom für diese Worte.
Aligned	Ich bitte darum sich etwas mehr Mühe bei der Argumentation zu geben, auch wenn man in der Mehrheit ist.
	Editdistance: 25 (96.2% WER)
Manual	Gemessen an den Gesamtausgaben ist das nicht viel.
Aligned	Gemessen an den Gesamtausgaben ist das nicht viel.
	Editdistance: 0 (0% WER)

Table 3: *Randomly selected examples of noisy labels from forced aligment of SPC-private audio recordings and transcripts. We observe added, missing and altered parts.*

4.4. Noisy Student Training

The noisy student experiments rely on the baseline described in Section 4.2 and use the same network architecture, training hyperparameters, language model and decoding parameters.

Data. The teacher model is trained on audios from the SPC-public corpus. The student model is trained on pseudo labeled SPC-private speech data. We use two versions of the SPC-private corpus. For the first version, one audio equals one sentence according to the forced alignment. The second version uses Voice Activity Detection (VAD) to split the audios into chunks of lengths between 2 and 12 seconds.

Noisy Student training. The teacher model is equal to the baseline. We create pseudo labels for the SPC-private dataset with the trained teacher model. This pseudo labeled data is used together with the SPC-public data to train the student model in the same way the teacher is trained. We train two model generations. The first generation is trained on the SPC-public corpus. Due to computational limitations, we were not able to train further generations.

4.5. Wav2vec 2.0

Wav2vec 2.0 models are implemented in FairSeq [25]. The model configuration used for the experiments in this work correspond to the BASE setup described in [3]. In this setup, the feature encoder consists of seven temporal convolutions with channel dimension 512 that reduce the temporal dimension by a factor of about 0.003. The subsequent Transformer network contains 12 transformer blocks with embedding dimension 512, inner dimension (FFN) 3072, and 8 attention heads.

Data. For the pretraining, we use the combined audios from the SPC-public corpus and the SPC-private corpus. As in noisy student training, we again use two versions of the SPC-private corpus. For the first version, we employ VAD to split the raw audios into smaller chunks of lengths between 2 and 36 seconds, and for the second version we take the splits from the forced alignment. For the finetuning, we use the SPC-public corpus without IoU filter.

Pretraining. We run the training for a total of 400K updates, with a learning rate of 0.0005. The learning rate is scheduled to warm up for the first 32K updates and then linearly decay for the rest of the training. We retain the checkpoint with the lowest loss computed on the validation split for finetuing.

Finetuning. For finetuning, we add a linear classification layer on top of the pretrained wav2vec model that maps representation vectors to output vocabulary. We train the model on the SPC-public corpus for 50K updates on the CTC loss with a learning rate of 0.003. Following Baevski et al. [3], we keep the pretrained weights constant for the first 10K updates and employ a learning rate schedule that consists of warmup (10%), constant (40%) and exponential decay (50%) stages.

Language Model. We use KenLM [26] to train 4-gram language models on the German sentences described in Section 2.

5. Results

We present the results of our experiments in Table 4. The supervised learning baseline model trained on SPC-public achieved a WER of 32.8%, measured on the SPC-public test set.

We observed the overall best performance by augmenting SPC-public with noisily labeled data from SPC-private. The resulting model achieved a WER of 26.4% on the SPC-public test set which means a relative improvement of about 20% compared to the baseline. This model was trained with five times more labeled data.

Using the additional data of SPC-private with noisy student training, we observed a relative WER improved by about 5% to 8% compared to the baseline. Thereby, using the audio split from forced alignment (Noisy Student (2)) yielded a slightly better performance compared to the VAD audio split (Noisy Student (1)).

Wav2Vec 2.0 pretraining on SPC-private followed by finetuning on SPC-public could improve the WER by about a relative of 15% compared to the baseline. This means that wav2vec 2.0 yielded better results than noisy student training, but fell short of supervised training on all data with noisy labels. We ob-

Model	LM	Model parameters	WER		BLEU	
			valid	test	valid	test
Baseline	LSTM	27M	32.6	32.8	49.5	49.7
Supervised Learning with Noisy Labels	LSTM	27M	**26.3**	**26.4**	**58.4**	**58.3**
Noisy Student (1)	LSTM	27M	30.9	31.1	52.1	51.9
Noisy Student (2)	LSTM	27M	30.5	30.3	53.5	53.9
Wav2vec (1)	4-gram	95M	28.3	27.9	55.0	55.5
Wav2vec (2)	4-gram	95M	28.3	27.8	54.8	55.4

Table 4: *Word error rate (WER) in % and BLEU score on the SPC-public validation and test set. Noisy student (1) and Wav2vec (1) indicate the experiments conducted with audio splits obtained from VAD. Noisy student (2) and Wav2vec (2) indicate the experiments conducted with audio splits obtained from forced alignment.*

served no significant differences between audio split from VAD (Wav2vec (1)) and those from forced alignment (Wav2vec (2)).

6. Conclusion

We compared three different methods for Swiss German ASR on data with noisy labels. We argue that the conclusions of this work can be transferred to low-resource tasks with similar characteristics.

We showed that even with imperfect transcriptions that result in low label quality, supervised learning on this data outperforms noisy student and wav2vec 2.0 pretraining on the same data. While this requires access to transcribed data, removing the constraint of exact transcriptions increases the amount of usable data. Additionally, one needs access to an existing ASR model to perform the forced alignment. Plüss et al. have shown that a relatively weak ASR model with a WER of 62% is sufficient to perform a forced alignment [7].

Park et al. have shown that training multiple model generations in the noisy student setting improves the WER considerably [8]. It remains unclear if the same effect holds in our setting. In future work, it would be interesting to see if training the initial teacher model on all available data with noisy labels and proceeding with the noisy student training improves the performance further. Also, investigating the impact of post-processing that improves the labels would be an interesting path to follow. For example, Plüss et al. have shown that filtering labels according to a tuned quality estimator leads to a WER improvement of 1% [7].

7. Acknowledgements

We would like to thank the parliamentary services of the cantons of Appenzell Ausserrhoden and Obwalden and the city of Bern for their work on the transcription of the debates and for sharing the recordings with us.

Calculations were performed at the sciCORE[2] scientific computing center at University of Basel.

This work was supported by Swiss National Science Foundation within the project "End-to-End Low-Resource Speech Translation for Swiss German Dialects (E2E_SG)" [205121_200729/1].

[2]http://scicore.unibas.ch/

8. References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, 2020.

[2] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, 2020.

[4] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[5] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[7] M. Plüss, L. Neukom, C. Scheller, and M. Vogel, "Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus," in *Proceedings of the Swiss Text Analytics Conference 2021*, 2021.

[8] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *Proc. Interspeech 2020*, 2020.

[9] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[10] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021.

[11] P. Garner, D. Imseng, and T. Meyer, "Automatic speech recognition and translation of a swiss german dialect: Walliserdeutsch," in *Proc. Interspeech 2014*, 2014.

[12] P. Dogan-Schönberger, J. Mäder, and T. Hofmann, "Swissdial: Parallel multidialectal corpus of spoken swiss german," *arXiv preprint arXiv:2104.02133*, 2021.

[13] M. Plüss, L. Neukom, and M. Vogel, "Swisstext 2021 task 3: Swiss german speech to standard german text," in *Proceedings of the Swiss Text Analytics Conference 2021*, 2021.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[15] M. Plüss, L. Neukom, and M. Vogel, "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)*, 2020.

[16] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *Proc. Interspeech 2020*, 2020.

[17] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[18] A. Köhn, F. Stegen, and T. Baumann, "Mining the spoken wikipedia for speech data and beyond," in *LREC*, 2016.

[19] P. Dogan-Schönberger, J. Mäder, and T. Hofmann, "Swissdial: Parallel multidialectal corpus of spoken swiss german," *arXiv preprint arXiv:2103.11401*, 2021.

[20] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *The Tenth Machine Translation Summit Proceedings of Conference*, 2005.

[21] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri, "Findings of the 2019 conference on machine translation (WMT19)," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 2019.

[22] C. Rauh and J. Schwalbach, "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies," 2020.

[23] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH*. ISCA, 2015.

[24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech (2018)*, 2018.

[25] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL-HLT (Demonstrations)*, 2019.

[26] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011.