



Large-Scale Self- and Semi-Supervised Learning for Speech Translation

Changhan Wang*, Anne Wu*, Juan Pino*, Alexei Baevski, Michael Auli, Alexis Conneau

Facebook AI, USA

{changhan, juancarabina, abaevski, michaelauli, aconneau}@fb.com, annewu018@gmail.com

Abstract

In this paper, we improve speech translation (ST) through effectively leveraging large quantities of unlabeled speech and text data in different and complementary ways. We explore both pretraining and self-training by using the large Libri-Light speech audio corpus and language modeling with Common-Crawl. Our experiments improve over the previous state of the art by 2.8 BLEU on average on all four considered CoVoST 2 language pairs via a simple recipe of combining wav2vec 2.0 pretraining, a single iteration of self-training and decoding with a language model. Different from existing work, our approach does not leverage any other supervision than ST data. Code and models are publicly released¹.

Index Terms: speech translation, unsupervised pretraining, self-training, semi-supervised learning

1. Introduction

Similar to many machine learning tasks, the amount of labeled data for speech-to-text applications such as automatic speech recognition (ASR) and speech translation (ST) is limited. For speech and language tasks, this problem is exacerbated by the fact that people speak many languages around the world and obtaining large quantities of labeled speech translation data for every language is simply not practical. On the other hand, unlabeled speech audio or text data is much more plentiful and various techniques to utilize it have been explored.

Semi-supervised learning techniques for ASR such as unsupervised speech pretraining [1, 2, 3, 4] and self-training or (iterative) pseudo-labeling [5, 6, 7, 8] have substantially improved performance on the traditional LibriSpeech benchmark, and led to systems that can learn with little supervision. In addition, it was recently shown that self-training and self-supervised pretraining can be effectively combined [9, 10] for speech recognition to achieve 4.8 WER on LibriSpeech test-other with only 10 minutes of annotated speech data.

Similarly for speech translation, there has been work on combating data scarcity, namely multitask learning [11, 12, 13], pretraining on ASR data [12, 14, 15, 16], data augmentation [17, 18, 19, 20], self-supervised pretraining [21, 22], self-training [23] or multilingual speech translation [24, 25, 26, 27, 28]. However, multitask learning, pretraining, data augmentation and multilingual ST rely on additional supervision provided by labeled ASR data, machine translation (MT) data or ST data while self-training and self-supervised pretraining exploit unlabeled source speech data. Unlike in ASR, the complementarity of self-supervised learning and self-training (and other semi-supervised techniques) has not been studied for the ST task.

On the other hand, past work in ST relied heavily on labeled data in the form of speech-to-text translation data, ASR

transcriptions [29] or MT sentence pairs to improve performance [28]. In this work, we follow the trend of leveraging purely unlabeled data to improve performance and show that this obtains strong performance while leveraging only supervision from ST data. Using wav2vec 2.0 pretraining, self-training and language model decoding, we show that we can outperform previous work while leveraging much less supervision.

Our contributions are as follows: we present a comprehensive study of the impact of existing semi-supervised learning techniques on speech translation and show that they greatly reduce the need for additional supervision in the form of labeled ASR or translation parallel data. We show that our simple approach obtains state-of-the-art results on all four language pairs we evaluate on: English to German, English to Catalan, English to Arabic and English to Turkish. We also conduct an ablation study on the impact of the quantity of unlabeled data for self-training and self-supervised pre-training in the context of ST.

In what follows, we describe the semi-supervised learning techniques and the system we use for ST. Then we present our results on the CoVoST 2 benchmark [27] on four language pairs and compare our work to the literature.

2. Learning from Unlabeled Data

In this section, we describe the techniques to leverage unlabeled speech or text data which we use in this study.

2.1. Unsupervised Pretraining

Unsupervised pretraining has been very effective in multiple fields of machine learning, including natural language processing [30], computer vision [31, 32] and speech recognition [3]. In this work, we demonstrate the impact of unsupervised pretraining for speech translation (ST) by leveraging a wav2vec 2.0 model pretrained on Libri-Light², a unlabeled dataset comprising 53K hours of English read audio books. The model is trained by predicting the latent speech representations of masked time-steps using a loss similar to SimCLR [32]. The latent speech representations are quantized for the prediction task and there is a fixed number of latents stored in a codebook.

For ST, we construct a sequence-to-sequence model with attention [33, 34] by adding a randomly initialized decoder model on top of a wav2vec 2.0 encoder. After pretraining, all parameters are fine-tuned on the CoVoST 2 ST data [27]. The decoder is also a Transformer model but smaller, with 7 layers and model dimension 256, which we do not pretrain³.

2.2. Self-Training

Self-training is a semi-supervised learning method that first trains a *teacher model* on labeled data. The teacher model is

²<https://github.com/pytorch/fairseq/blob/master/examples/wav2vec>

³Pretraining with a masked language model did not improve performance in our setting

* Equal contribution.

¹https://github.com/pytorch/fairseq/tree/master/examples/speech_to_text

Table 1: *CoVoST 2 dataset*. We use four language pairs of the CoVoST 2 dataset for our experiments, all from English due to the easy access of open-source English unannotated data in that source language.

Language pair	en-de	en-ca	en-ar	en-tr
Speech hours	430h	430h	430h	430h
Target utterances	288K	288K	288K	288K
Target words	2.8M	3.1M	2.4M	2.2M

then used to synthetically annotate unlabeled data in order to train a new *student model* on the combination of labeled and pseudo-labeled data [35, 23, 36]. We further fine-tune the student model on labeled data to alleviate the domain mismatch between labeled and unlabeled data. Recent work showed that unsupervised pretraining and self-training can be complementary for natural language understanding [37] and speech recognition [9]. In this work, we adopt this setting by considering self-training on top of a wav2vec 2.0 pretrained model and show the complementarity of these two learning approaches for speech translation.

2.3. Decoding with Language Model

With unsupervised pretraining and self-training, we leverage additional unlabeled speech data to improve the performance of a ST system. We also make use of monolingual text data in the target language to further improve translation quality. Specifically, we train a language model (LM) on part of the CommonCrawl data⁴ that is in similar domains as CoVoST 2 (§ 3.1). And then we combine ST model and LM scores at every time step in beam search decoding (shallow fusion) [38].

3. Datasets and Training Details

In this section, we describe the datasets we use for speech translation (ST), the setups for unsupervised pretraining, self-training as well as language modeling. Then we give details on how we train our models.

3.1. Datasets

CoVoST 2: speech translation data. CoVoST 2⁵ is a large-scale multilingual speech translation corpus covering translations from 21 languages into English and from English into 15 languages. This represents the largest open dataset available to date from total volume and language coverage perspective. Specifically, we cover four language directions from English to German (de), Catalan (ca), Arabic (ar) and Turkish (tr), which contain 430 hours of annotated data each. For simplicity, we choose to focus on speech translation from English because of easy public access to unlabeled data in that language.

Libri-Light: pretraining and self-training data. We use the wav2vec 2.0 model⁶ pretrained on the Libri-Light data [39], a dataset consisting of more than 60k hours of unlabeled read speech data⁷. The dataset is derived from

open-source audio books from the LibriVox project and is the largest freely-available corpus of speech. Previous work in unsupervised pretraining showed that using Libri-Light over LibriSpeech led to better performance [3]. We use the 6000-hour subset of Libri-light (LV-6k) as well as the full 60k hours dataset (LV-60k) for self-training. Specifically, we train models on CoVoST 2 whose domain is read speech, and synthetically label unannotated data from LV-60k - which is also read speech. We then fine-tune a wav2vec 2.0 pretrained model - the student model - on both the synthetically annotated data and the ground-truth CoVoST 2 training data. Self-training is a good alternative to back-translation in the case of speech translation, although the target data generated by the teacher model is not real data. To remedy this issue, we leverage ground-truth target data through language model decoding. Note that similar to NoisyStudent, we also inject noise in our input during self-training through the same masking strategy as wav2vec 2.0 [40, 3].

LibriSpeech: pretraining and self-training ablation data. In order to study the effect of the amount of unlabeled data on both pretraining and self-training, we use the 960 hours of LibriSpeech as a smaller-scale alternative to Libri-light and LV-60k for wav2vec 2.0 pretraining⁸ and for self-training.

CommonCrawl: language model data. In order to obtain language model data in the right domain, we leverage CommonCrawl (CC) data from the CC100 corpus [42, 43] for the four target languages studied (German, Catalan, Arabic, Turkish). First, we train 4-gram language models with the KenLM toolkit [44] on the training set of the CoVoST 2 data and on the CC data and filter CC by averaging the LM scores [45]. We then keep only one tenth of the original data, and use transformer-based language models trained on both the CoVoST 2 training set and the additional CC data for language model rescoring.

3.2. Training Details

Unsupervised pretraining. For wav2vec 2.0 pretraining, we use the *Large* model, which comprises 24 self-attention blocks with model dimension 1024, inner dimension 4096 and 16 attention heads, resulting in a total of about 300M parameters. The feature encoder contains seven blocks and the temporal convolutions in each block have 512 channels with strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,3,3,2,2), resulting in a receptive field of about 25ms and a stride of about 20ms.

Speech translation. We use a sequence-to-sequence model where the encoder is a wav2vec 2.0 model with several layers of convolutions followed by a Transformer network with 24 layers. The decoder is also a Transformer network with 7 layers, an embedding size of 256, 4 attention heads and feed forward network (FFN) dimension of 2048. A 10K byte-pair encoding (BPE) vocabulary is built on the CoVoST 2 target text for each target language. We train our model with Adaptive Moment estimation (Adam) [46], a learning rate of 3e-4, label smoothing with probability 0.1, an effective batch size of 6.4M tokens, layer drop 0.05, a masking strategy similar to wav2vec 2.0 with mask length 5 and mask probability 0.15. During the fine-tuning phase, the wav2vec 2.0 encoder is frozen for 10K updates. Models are trained for 250K updates and the best checkpoint is selected based on the BLEU score on the validation set. The pseudo-labels are generated with a beam size of

⁴<http://data.statmt.org/cc-100/>

⁵<https://github.com/facebookresearch/covost>

⁶dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_vox_new.pt

⁷<https://github.com/facebookresearch/libri-light>

⁸dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt

Table 2: BLEU on four language pairs of CoVoST-V2: English-German (*en-de*), English-Catalan (*en-ca*), English-Arabic (*en-ar*) and English-Turkish (*en-tr*). The results show that self-supervised pre-training on Librispeech (*wav2vec-2.0*) followed by self-training on the same data can improve performance. Using a language model during decoding improves performance further.

Row	Model	en-de	en-ca	en-ar	en-tr	Avg
<i>Baselines from CoVoST 2</i>						
1	Wang et al. [27] (w/o pre-ASR)	13.6	20.2	8.7	8.9	12.9
2	Wang et al. [27] (w/ pre-ASR)	16.3	21.8	12.1	10.0	15.1
<i>Previous state-of-the-art results</i>						
3	Prev E2E SOTA [41, 27, 28]	18.4	23.6	13.9	11.7	16.9
4	Cascade SOTA [28]	19.4	25.0	14.3	11.7	17.6
5	Li et al. [28] (joint training)	25.8	30.9	18.0	17.0	22.9
6	Li et al. [28] (+ extra MT data)	26.6	30.4	18.6	16.3	23.0
<i>Our results</i>						
8	wav2vec-2.0	25.5	33.1	19.9	17.4	24.0
9	wav2vec-2.0 + decoding w/ LM	25.9	33.4	20.4	17.7	24.4
10	wav2vec-2.0 + self-training (LV-60k)	27.1	34.4	21.3	18.8	25.4
11	wav2vec-2.0 + self-training (LV-60k) + decoding w/ LM	27.7	34.8	21.6	19.2	25.8

4.

Language models. For the language models, we use the Google Billion Word Transformer architecture of [47] with 12 decoder layers and an embedding dimension of 512. We train the model with Adam, and an inverse sqrt scheduler. When using LM scores in decoding, we scale the scores by 0.1 and use a length penalty of 0.7.

4. Experiments and Results

In this section, we describe our experimental results obtained in Table 2 and 3 where we combine self-supervised and semi-supervised learning techniques.

4.1. Improvements from Unsupervised Pretraining

We observe strong gains using wav2vec 2.0 models compared to previous baselines which were using similar speech translation architectures without pretraining. The Libri-light pretrained wav2vec 2.0 model (row 8) achieves 24.0 BLEU on average, which is on average 11.1 BLEU points better than the baseline of [27] (row 1) and 8.9 BLEU higher than their model which leveraged ASR pretraining as additional supervision (row 2). These results demonstrate that the features learned by the wav2vec 2.0 model are very useful beyond speech recognition, and applicable to other speech tasks such as speech translation. The results with only wav2vec 2.0 pretraining (row 8) outperform the most recent state-of-the-art results (row 6) on CoVoST 2 [28] for three out of the four language pairs (except English-German), without using any other supervision than the speech translation data.

4.2. Improvements from Self-Training

As previously shown for computer vision [48], natural language understanding [37] and speech recognition [9], supervised or unsupervised pretraining can be complementary to self-training. Combining self-supervised learning with semi-supervised learning for speech translation in this work, we first use the previously described Libri-Light wav2vec 2.0 models fine-tuned on the CoVoST 2 speech translation data as teacher

models, and synthetically annotate the 60k-hour Libri-light dataset (LV-60k). We then leverage self-training by finetuning a wav2vec 2.0 pretrained student model on both LV-60k synthetic data and the CoVoST data [49, 23]. Note that we upsample the ground-truth CoVoST data such that it has the same importance during finetuning than the synthetically annotated LV-60k. After following this procedure, we obtain 25.4 average BLEU score (row 10) which is 1.4 BLEU better than using wav2vec 2.0 pretraining only. This demonstrates the complementarity of pretraining and self-training for the task of speech translation. The improvement is consistent across language pairs. With this method, we reach a new state of the art on the CoVoST 2 benchmark for all four considered language pairs (row 10).

4.3. Improvements from Decoding with Language Model

Self-training and unsupervised pretraining both leverage additional unannotated speech data to improve performance. But self-training generates noisy output on which the student model is fine-tuned which may lead the model to learn incorrect patterns. To inject more prior knowledge about the target language structure, a natural solution is to use unannotated text in the target domain. In this work, we leverage language modeling as one way to do that, and use it to improve generation through decoding. This improves the wav2vec 2.0 baseline by 0.4 BLEU on average across all language pairs (row 9 vs. row 8). It also improves the stronger setting of wav2vec 2.0 + self-training by 0.4 BLEU (row 11 vs. row 10). Combining wav2vec 2.0 pretraining, self-training and language model decoding in row 11, we reach a new state of the art on the CoVoST 2 benchmark, with an average BLEU score of 25.8 over the four language pairs.

4.4. Comparison with Previous Work

The combination of pretraining, self-training and LM decoding outperforms the prior state of the art [28] in all language directions, and by 2.8 BLEU on average. The prior state of the art uses both pretraining with wav2vec 2.0 and mBART, as well as a minimalistic LNA (LayerNorm and Attention) fine-tuning. We note that mBART was fine-tuned using additional

Table 3: Effect of the amount of unlabeled data on pretraining and self-training (cf. Table 2).

Row	Model	en-de	en-ca	en-ar	en-tr	Avg
<i>Increasing the amount of pretraining data</i>						
1	wav2vec-2.0 (LS-960)	20.5	27.3	15.4	14.0	19.3
2	wav2vec-2.0 (LV-60k)	25.5	33.1	19.9	17.4	24.0
<i>Increasing the amount of self-training data</i>						
3	wav2vec-2.0 + Self-training (LS-960)	25.7	32.9	20.0	17.6	24.1
4	wav2vec-2.0 + Self-training (LV-6k)	27.0	34.1	21.4	18.4	25.2
5	wav2vec-2.0 + Self-training (LV-60k)	27.1	34.4	21.3	18.8	25.4

labeled machine translation data and therefore our approach relies on less labeled data. Moreover, this work is conducted on a bilingual setting without additional supervision coming from multiple language pairs. Finally, the prior state-of-the-art model uses an adapter between the encoder and the decoder that further downsamples the input by a factor of 8. In contrast, our model architecture was simplified by removing the adapter module, as well as the LNA finetuning.

4.5. Data Ablation for Unsupervised Pretraining and Self-Training

Prior work has shown that increasing the amount of unlabeled data for pretraining and self-training can improve ASR performance [3, 9]. To understand whether the same holds true for ST, we compare pretraining a wav2vec 2.0 model on the 960 hours unlabeled speech audio from the smaller LibriSpeech corpus (LS-960) instead of the 60k hours of the LibriVox corpus (LV-60k) used so far. Table 3 shows that pretraining the speech encoder on more data leads to a large improvement of 4.7 BLEU on average across the four language pairs (row 2 vs. row 1). Note that pretraining wav2vec 2.0 on LibriSpeech still provides an average improvement of 4.2 BLEU over the supervised baseline with ASR pretraining (row 1 vs. row 2 from Table 2).

Next, we examine increasing the amount of unlabeled data for self-training. We compare pseudo-labeling LS-960 with pseudo-labeling LV-6k and LV-60k. From Table 3 we can see that the increase of self-training data from LS-960 to LV-6k brings an additional gain of 1.1 BLEU on average (row 4 vs. row 3). Scaling self-training even further to LV-60k leads to an additional gain of 0.2 BLEU on average (row 5 vs. row 4). For both pretraining and self-training, we found that gains using more unlabeled data were similar when using a language model.

5. Conclusion

We pushed the limits of self-supervised and semi-supervised learning for speech translation by leveraging pretraining with wav2vec 2.0 and self-training. These techniques can outperform the previous state of the art by an average of 2.8 BLEU across four language directions without using any type of supervision other than the CoVoST 2 data. We also demonstrated the complementarity of unsupervised pretraining, self-training and language model decoding. Our work provides stronger and simpler baselines for speech translation and demonstrates the effectiveness of wav2vec 2.0 unsupervised pretraining for speech translation.

6. References

- [1] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. abs/1807.03748, 2018.
- [2] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019.
- [3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.
- [4] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv*, vol. abs/2006.13979, 2020.
- [5] G. S. et al., "End-to-end ASR: from supervised to semi-supervised learning with modern architectures," *arXiv*, vol. abs/1911.08460, 2019.
- [6] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *Proc. ICASSP*, 2020.
- [7] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *arXiv*, vol. abs/2005.09267, 2020.
- [8] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimipl: Language-model-free iterative pseudo-labeling," *arXiv*, vol. abs/2010.11524, 2020.
- [9] 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 *Proc. ICASSP*, 2020.
- [10] 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*, vol. abs/2010.10504, 2020.
- [11] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *arXiv preprint arXiv:1703.08581*, 2017.
- [12] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proc. ICASSP*, 2018.
- [13] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *Proc. ICASSP*, 2021.
- [14] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proc. NAACL*, 2019.
- [15] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pre-training for low-resource speech-to-text translation," in *Proc. ICASSP*, 2020.
- [16] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *Proc. AAAI*, 2020.

- [17] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *Proc. ICASSP*, 2019.
- [18] J. Pino, L. Puzon, J. Gu, X. Ma, A. D. McCarthy, and D. Gopinath, "Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade," in *Proc. IWSLT*, 2019.
- [19] E. Salesky, M. Sperber, and A. W. Black, "Exploring phoneme-level speech representations for end-to-end speech translation," in *Proc. ACL*, 2019.
- [20] A. D. McCarthy, L. Puzon, and J. Pino, "Skinaugment: Auto-encoding speaker conversions for automatic speech translation," in *Proc. ICASSP*, 2020.
- [21] A. Wu, C. Wang, J. Pino, and J. Gu, "Self-Supervised Representations Improve End-to-End Speech Translation," in *Proc. Interspeech*, 2020.
- [22] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève, and L. Besacier, "Investigating Self-Supervised Pre-Training for End-to-End Speech Translation," in *Proc. Interspeech*, 2020.
- [23] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-Training for End-to-End Speech Translation," in *Proc. Interspeech*, 2020.
- [24] M. A. Di Gangi, M. Negri, and M. Turchi, "One-to-many multilingual end-to-end speech translation," in *Proc. ASRU*, 2019.
- [25] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual end-to-end speech translation," in *Proc. ASRU*, 2019.
- [26] C. Wang, J. Pino, A. Wu, and J. Gu, "Covost: A diverse multilingual speech-to-text translation corpus," in *Proc. LREC*, 2020.
- [27] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv*, 2020.
- [28] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv*, vol. abs/2010.12829, 2021.
- [29] S. B. H. K. K. Livescu and A. L. S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proc. NAACL*, 2019.
- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019.
- [31] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv*, vol. abs/1911.05722, 2019.
- [32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv*, vol. abs/2002.05709, 2020.
- [33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Proc. NIPS*, 2014.
- [34] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015.
- [35] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, and et al., "Improved noisy student training for automatic speech recognition," *arXiv*, 2020.
- [36] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *Proc. ICASSP*, 2020.
- [37] J. Du, E. Grave, B. Gunel, V. Chaudhary, O. Celebi, M. Auli, V. Stoyanov, and A. Conneau, "Self-training improves pre-training for natural language understanding," *arXiv*, vol. abs/2010.02194, 2020.
- [38] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.
- [39] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, and et al., "Libri-light: A benchmark for asr with limited or no supervision," *arXiv*, vol. abs/1912.07875, 2019.
- [40] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.
- [41] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, and et al., "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *Proc. ICASSP*, 2020.
- [42] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. ACL*, 2020.
- [43] G. Wenzek and et al., "CCNet: Extracting high quality monolingual datasets from web crawl data," in *Proc. LREC*, 2020.
- [44] K. Heafield, I. Pouzyrevsky, and et al., "Scalable modified Kneser-Ney language model estimation," in *Proc. ACL*, 2013.
- [45] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," *Proc. ACL*, 2010.
- [46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [47] A. Baevski and M. Auli, "Adaptive input representations for neural language modeling," *arXiv preprint arXiv:1809.10853*, 2018.
- [48] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, and et al., "Rethinking pre-training and self-training," *arXiv*, vol. abs/2006.06882, 2020.
- [49] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," in *ICLR*, 2020.