Parameter-Efficient Transfer Learning for End-to-end Speech Translation

Yunlong Zhao, Kexin Wang, Qianqian Dong*, **Tom Ko**
ByteDance, China
{zhaoyunlong.123, wkx, dongqianqian, tom.ko}@bytedance.com

Abstract

Recently, end-to-end speech translation (ST) has gained significant attention in research, but its progress is hindered by the limited availability of labeled data. To overcome this challenge, leveraging pre-trained models for knowledge transfer in ST has emerged as a promising direction. In this paper, we propose PETL-ST, which investigates parameter-efficient transfer learning for end-to-end speech translation. Our method utilizes two lightweight adaptation techniques, namely prefix and adapter, to modulate Attention and the Feed-Forward Network, respectively, while preserving the capabilities of pre-trained models. We conduct experiments on MuST-C En-{De, Es, Fr, Ru} datasets to evaluate the performance of our approach. The results demonstrate that PETL-ST outperforms strong baselines, achieving superior translation quality with high parameter efficiency. Moreover, our method exhibits remarkable data efficiency and significantly improves performance in low-resource settings.

Keywords: Speech Translation, Transfer Learning, Low Resource

1. Introduction

The end-to-end speech-to-text translation (ST) has garnered significant attention (Wang et al., 2020a,c; Dong et al., 2021a,b) due to its ability to alleviate error propagation and reduce latency compared to cascade systems (Sperber and Paulik, 2020; Lam et al., 2021; Bentivogli et al., 2021). However, the availability of large-scale paired speech and translated text data is limited. This scarcity of data makes training end-to-end ST models from scratch challenging. Consequently, leveraging pre-trained models for end-to-end ST shows greater promise compared to training from scratch (Xu et al., 2021; Ye et al., 2021; Fang et al., 2022).

Several existing approaches have investigated the utilization of pre-trained models for speech translation, employing techniques such as multi-task learning (Ye et al., 2021; Han et al., 2021; Fang et al., 2022) and knowledge distillation (Liu et al., 2019; Tang et al., 2021; Inaguma et al., 2021). While these methods have demonstrated impressive performance improvements in ST, there are still underlying issues that require attention. Firstly, fine-tuning pre-trained models for the ST task can disrupt their original parameters, potentially affecting the knowledge acquired during their pre-training phase. Secondly, since pre-trained models often have a significant number of parameters, fine-tuning all of them incurs high training and storage costs. Therefore, it is crucial to explore parameter-efficient transfer learning methods for the ST task.

Parameter-efficient methods have attracted much interest, primarily driven by the success of large pre-trained models (Brown et al., 2020; Liu et al., 2020). Previous efficient fine-tuning methods for ST can be categorized into two types: fine-tuning a subset of the model parameters or incorporating additional tunable modules. Under the first category, LNA (Li et al., 2021) focuses on fine-tuning only LayerNorm and Attention, demonstrating cross-modality transfer capability in multilingual speech translation scenarios. In the second category, Adapter modules are integrated to modify the output of the Feed-Forward Network (FFN) sub-layer in the Transformer for multilingual ST (Le et al., 2021).

In our work, we propose PETL-ST, a parameter-efficient transfer learning method for end-to-end ST. PETL-ST leverages the pre-trained acoustic and machine translation (MT) models, and adapts each sub-layer in the Transformer to facilitate speech-to-text translation. Drawing inspiration from the success of prefix-tuning (Li and Liang, 2021), we propose applying prefix-tuning to adapt the *Attention* module in the Transformer, rather than directly fine-tuning it. Additionally, we introduce the parallel adapter to adapt the *FFN* sub-layer and perform direct fine-tuning of *LayerNorm*. Our PETL-ST surpasses strong baselines and achieves comparable performance to full fine-tuning while utilizing fewer than 10% of the parameters. Furthermore, our research demonstrates outstanding data efficiency in the speech translation task.

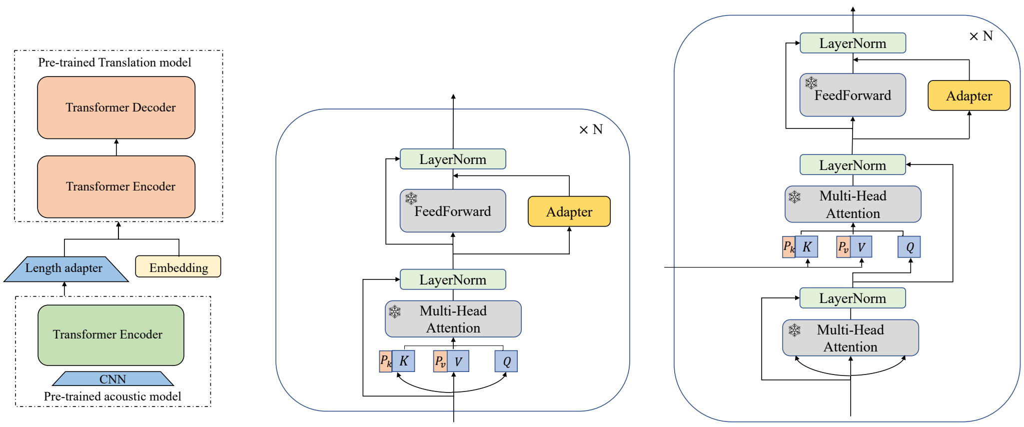
2. Methodology

2.1. Problem Formulation

End-to-end ST aims to directly translate speech signal sequence $\mathbf{s} = (s_1, \dots, s_{|\mathbf{s}|})$ in source language

*Corresponding author.

LREC-COLING 2024, pages 12592–12598



(a) Model Structure (b) Encoder (c) Decoder

Figure 1: (a) Overview of model structure. Our model comprises a pre-trained NMT model and a pre-trained acoustic model, wav2vec 2.0, with a CNN length adapter. (b) The encoder of PETL-ST. We add the prefix for Self-Attention and the parallel adapter for *FFN* in Encoder. (c) The decoder of PETL-ST. We add the prefix for Cross-Attention and the parallel adapter for *FFN* in Decoder.

to the text $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|})$ in another language. ST corpus usually has transcripts of the source language speech $\mathbf{x} = (x_1, \ldots, x_{|\mathbf{x}|})$, which can be denoted as $\mathcal{D}_{\mathrm{ST}} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$. And MT corpus can be denoted as $\mathcal{D}_{\mathrm{MT}} = \{(\mathbf{x}, \mathbf{y})\}$.

2.2. Model Architecture

As Figure 1(a) illustrates, our model has a pre-trained MT model and a pre-trained acoustic model. We keep the parameters of each pre-trained model frozen and fine-tune on the downstream ST task.

Pre-trained MT model We introduce a large Transformer-based (Vaswani et al., 2017) MT model for our framework, which is pre-trained on external datasets. The detailed configurations can be seen in Section 3.2.

Pre-trained acoustic model We use Wav2Vec2 (Baevski et al., 2020), which is pre-trained on large amounts of unlabeled audio data in a self-supervised manner. The acoustic model extracts high-level representations from the speech sequences $\mathbf{s}$ as the audio input of our framework, which can accept bi-modal inputs.

Length adapter We employ a convolutional sub-sampler, which consists of a stack of two 1-dimensional convolution layers, as the length adapter. As the length adapter is not included in the pre-trained models, its parameters are tunable and can be adjusted during fine-tuning.

2.3. Multi-task Objective

Multi-task learning has been proven to improve speech translation performance (Ye et al., 2021; Han et al., 2021). Thus, we train MT and ST tasks jointly. The corresponding speech transcripts are also used as input for MT training. The training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ST}} + \mathcal{L}_{\mathrm{MT}} \tag{1}$$

where

$$\begin{aligned} \mathcal{L}_{\mathrm{ST}} &= -\mathbb{E}_{\mathbf{s}, \mathbf{y} \in \mathcal{D}_{\mathrm{ST}}} \log P(\mathbf{y} \mid \mathbf{s}) \\ \mathcal{L}_{\mathrm{MT}} &= -\mathbb{E}_{\mathbf{x}, \mathbf{y} \in \mathcal{D}_{\mathrm{MT}}} \log P(\mathbf{y} \mid \mathbf{x}) \end{aligned} \tag{2}$$

2.4. Efficient Tuning

The Transformer (Vaswani et al., 2017) block is the main component of our model. A Transformer block consists of three main components, Attention, Feed-Forward Network, and LayerNorm. We design a parameter-efficient transfer learning framework to adapt each component for the speech translation task instead of full fine-tuning.

2.4.1. Prefix-tuning for Attention

Instead of fine-tuning all parameters in Attention, we use prefix-tuning for parameter-efficient transfer learning. Prefix-tuning is a lightweight tuning method inspired by prompt tuning (Lester et al., 2021), and it achieves comparable performance in many generation tasks (Tan et al., 2022). Specifically, the prefix, a sequence of continuous vectors, prepends to the original keys K and values V of the multi-head Attention at every layer. The prefix vectors, P_K and P_V, are tunable parameters to modulate the frozen pre-trained model. The formula for a single attention head is as follows:

$$\begin{aligned} \text{head}_{\text{i}} &= \operatorname{Attn}\left(QW_i^Q, \hat{K}W_i^K, \hat{V}W_i^V\right) \\ \hat{K} &= \operatorname{Concat}(P_K, K) \\ \hat{V} &= \operatorname{Concat}(P_V, V) \end{aligned} \tag{3}$$

Since our model is trained with multimodal inputs, we apply modality-aware prefix on the self-attention of the encoder (including Wav2Vec2.0 encoder and the pre-trained MT encoder). In the audio encoder (Wav2Vec2.0), the audio prefix prepends to the audio input representation. In the text encoder, the text prefix prepends to the text input representation. We also apply prefix-tuning on the cross-attention of the decoder. The prefix for cross-attention prepended to encoder outputs can unify encoder outputs for translation decoding. Figure 1(b) and 1(c) illustrate the design of PETL-ST for encoder and decoder.

2.4.2. Parallel Adapter-tuning for FFN

As for *FFN*, we apply adapter-tuning for parameter-efficient transfer learning. We follow the design by Houlsby et al. (2019), of which the adapter consists of two full-connected layers (down-projection $\boldsymbol{W}_{\text{down}}$ and up-projection $\boldsymbol{W}_{\text{up}}$) with a nonlinear activation function. And He et al. (2021) has shown that a parallel adapter is the best option to modulate *FFN*. Thus, we add the adapter for *FFN* in a parallel manner as:

$$\begin{aligned} \boldsymbol{h}_{\text{adapter}} &= \boldsymbol{W}_{\text{up}} \operatorname{ReLU}\left(\boldsymbol{W}_{\text{down}} \boldsymbol{h}\right) \\ \hat{\boldsymbol{h}} &= \boldsymbol{h}_{\text{adapter}} + \mathit{FFN}\left(\boldsymbol{h}\right) \end{aligned} \tag{4}$$

where $\boldsymbol{h}$ denotes the input of *FFN*, and $\hat{\boldsymbol{h}}$ denotes the “adapted” output.

2.4.3. LayerNorm

LayerNorm contains very few parameters trained based on the statistics of the data used in pretraining. It’s necessary to finetune these parameters during transfer learning for speech translation. Thus we make *LayerNorm* parameters tunable.

3. Experiments

3.1. Datasets

ST datasets We conduct our experiments on MuST-C, a multilingual speech translation dataset. MuST-C (Di Gangi et al., 2019) contains eight translation directions from English to different target languages. We focus on four translation directions language: EN→DE (German), ES (Spanish), FR (French), and RU (Russian).

MT Datasets To obtain well-trained pre-trained MT models for downstream ST, we first pre-train an MT model on an external MT dataset for each translation direction. In our experiments, we use WMT (Bojar et al., 2016) as the external datasets for pre-training MT models following Fang et al. (2022). Table 1 shows detailed statistics of all datasets.

Table 1: Statistics of all datasets.

	ST (MuST-C)		MT	
En →	hours	#sents	name	#sents
De	408	234K	WMT16	4.6M
Fr	492	280K	WMT14	40.8M
Ru	489	270K	WMT16	2.5M
Es	504	270K	WMT13	15.2M

3.2. Experimental Setups

Model Configuration For the pre-trained acoustic model Wav2vec2.0, we use the base configuration in (Baevski et al., 2020), which is only pre-trained on Librispeech without ASR finetuning [1]. For the pre-trained MT model, we use the large configuration transformer-based MT model. We also conduct experiments with the base configuration transformer-based MT model and discuss the effect of model capacity in the analysis section. The base configuration transformer MT model has $N_e = 6$ encoder layers and $N_d = 6$ decoder layers.

The detail settings of parameters-efficient tuning are as follows: As for prefix-tuning, since the length adapter is a 4× downsampling, the length of prefix vectors prepend to Wav2vec2.0, and the MT model is 4:1. Our experiments use the prefix length $|P_{mt}| = 50$ for MT model, as well $|P_{wav}| = 200$ for Wav2vec2.0 transformer encoder. Following previous work, we use a large feed-forward neural network to reparametrize the prefix vectors. As for adapter tuning, the bottleneck of two full-connected layers has 256 hidden units. We implement our models in Fairseq (Ott et al., 2019).

Experimental Configuration In MT pretraining, the learning rate is 7e-4, and each batch consists of 33k input tokens. In fine-tuning, the learning rate is 1e-3, and each batch comprises 16M source audio frames. The dropout is set to 0.1, and the value of label smoothing is 0.1. For the full fine-tuning as the comparison, the learning rate is set to 1e-4, and other settings are the same as parameter-efficient fine-tuning. We use beam search for decoding with a beam size of 5. We compute case-sensitive detokenized BLEU (Papineni et al., 2002) for evaluation using sacreBLEU[2] (Post, 2018).

4. Results and Analysis

4.1. Results on MuST-C Dataset

Table 2 shows the results on the MuST-C `tst-COMMON` set. Our method surpasses the strong

[1] `https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt`

[2] `https://github.com/mjpost/sacrebleu`, BLEU Signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0

Table 2: Case-sensitive detokenized BLEU scores on MuST-C `tst-COMMON` set. #Params. denotes the number of tunable parameters during training.

Models	#Params.	BLEU			
		De	Es	Fr	Ru
W2V2-Transformer	156M	26.9	30.0	36.6	17.3
LNA	138M	27.7	**31.7**	38.4	17.9
PETL-ST	35M	**27.9**	**31.7**	**38.7**	**18.0**

Table 3: Ablation results on MuST-C En-De `tst-COMMON` set. "#Params." denotes the number of tunable parameters during training.

Models	#Params.	BLEU (En-De)
PETL-ST	35M	**27.85**
w/o LayerNorm		27.73
Prefix-tuning-ST	17M	26.90
w/o LayerNorm		26.51
Adapter-tuning-ST	27M	27.57
w/o LayerNorm		27.54
Full Fine-tune	477M	27.82

baseline W2V2-Transformer (Fang et al., 2022) with only 20% trained parameters. The performance gains are proportional to the amount of pre-training MT data. It shows that our method can leverage MT efficiently for modality transfer learning and achieve comparable results in the speech translation task. As for the comparison with LNA fine-tuning (Li et al., 2021), our approach is more parametric efficient and better leverages the pre-trained model achieving better performance.

4.2. Ablation Studies

To explore the effect of each component, we conduct ablation studies on En-De direction. We refer to PETL-ST without adapter as *prefix-tuning-ST* and PETL-ST without prefix as *adapter-tuning-ST*. As shown in Table 3, *prefix-tuning-ST* achieves 95% performance with 3.6% trained parameters. And *adapter-tuning-ST* almost matches the results of full fine-tuning with only 5.6% trained parameters. And *LayerNorm* shows significance in our proposed method. Our method PETL-ST achieves a slight performance gain compared with full fine-tuning, which may be attributed to the over-fitting caused by the full fine-tuning of a large transformer.

4.3. Effect of Model Capacity

We conducted experiments to analyze the impact of model capacity. We apply PETL-ST approach to both base and large configuration MT models. The results are presented in Figure 2. PETL-ST applied to the base configuration did not achieve the same level of performance as full fine-tuning, which suggests that a larger model capacity plays a crucial role in achieving parameter-efficient tuning.

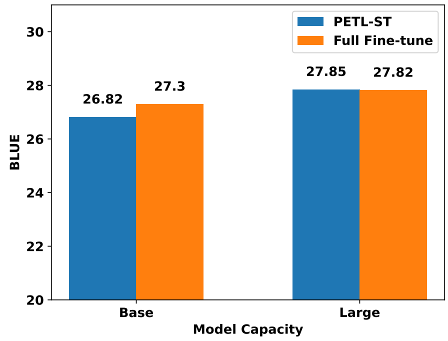


Figure 2: Effect of Model Capacity. Base and large refer to configurations of the pre-trained MT model.

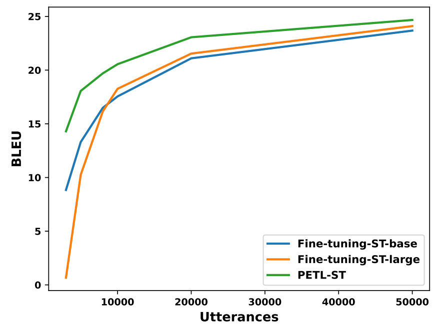


Figure 3: Comparison of Full fine-tuning and PETL-ST in low data resource setting. Base and large refer to configurations of the pre-trained MT model.

4.4. Low-resource Setting

To evaluate the data efficiency of PETL-ST, we conduct experiments in a low-resource setting. MuST-C En-De contains 408 hours of data, with 234K manual transcriptions and translations at the sentence level. We sample 3k to 50k utterances as the train set to simulate the low-resource setting and use the original `dev` set, and `tst-COMMON` set for evaluation. In Figure 3, we compare PETL-ST with full fine-tuning strategies in base and large model configuration. Our PETL-ST shows excellent data efficiency and the relative performance gain becomes more significant as the data resources decrease. In extremely low resources setting (3k utterances, less 2% of the total), full fine-tuning of a large model fails to train, while PETL-ST still achieves acceptable performance (14.30 BLEU). PETL-ST is an effective approach to exploiting large pre-trained models, especially in highly low-resource settings.

5. Conclusion

PETL-ST is an efficient fine-tuning method for ST. The proposed method leverages trainable prefixes and adapters to modulate sub-layers in transformer while preserving the original parameters of the pre-trained models as much as possible. Through experiments and analysis, we demonstrate that PETL-ST achieves promising performance, including remarkable data efficiency, by harnessing the powerful capabilities of pre-trained models.

6. Ethics Statement

Speech translation is an essential task of spoken language processing. This research paper on speech translation adheres to the highest ethical standards and considers the ethical implications and responsibilities. All speech and text data collected for this study were anonymized and treated with strict confidentiality. Personal identifying information such as names or contact details were removed or replaced with pseudonyms to ensure participant anonymity. We do not see any significant ethical concerns.

7. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021a. Consecutive decoding for speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12738–12748.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021b. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12749–12759.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1872–1881.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2021. Cascaded models with cyclic feedback for direct speech translation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7508–7512. IEEE.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer.

2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *ACL/IJCNLP (1)*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. *Interspeech 2019*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. 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)*, pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. Msp: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261.

Ioannis Tsiamas, Gerard I Gállego, Carlos Escolano, José Fonollosa, and Marta R Costa-jussà. 2022. Pretrained speech encoders and efficient fine-tuning methods for speech translation: Upc at iwslt 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 265–276.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020b. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proc. of AACL*.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020c. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *ACL (1)*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 2619–2630.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113.

Jinming Zhao, Hao Yang, Ehsan Shareghi, and Gholamreza Haffari. 2022. M-adapter: Modality adaptation for end-to-end speech-to-text translation. *arXiv preprint arXiv:2207.00952*.

8. Language Resource References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. of NAACL*.