CPT: CROSS-MODAL PREFIX-TUNING FOR SPEECH-TO-TEXT TRANSLATION

Yukun Ma, Trung Hieu Nguyen, Bin Ma

Alibaba Group
{yukun.ma,trunghieu.nguyen,bin.ma}@alibaba-inc.com

ABSTRACT

Speech translation models benefit from adapting multilingual pretrained language models. However, such adaptation modifies the parameters in the pretrained model to favor a specific task. Prefix-tuning, as a lightweight adaptation technique, has recently emerged as an efficient adaptation method that significantly reduces the number of trainable parameters and has demonstrated great potential in low-resource settings. It inserts prefixes into the output of each layer of a pretrained model, without modifying its parameters. During training, only the parameters of prefixes are updated while the rest of the model are being frozen. In this paper, we improve the performance of speech translation in medium-/low-resource settings by a cross-modal prefix that bridges the gap between speech input and translation modules to reduce the information loss in the cascaded model. We show that the proposed cross-modal prefix-tuning is effective, robust and parameter-efficient for adapting a speech recognition and translation pipeline.

***Index Terms*—** speech-to-text translation, prefix-tuning, lightweight adaptation

1. INTRODUCTION

Speech-to-text translation (ST) is concerned with translating human speech of a source language to the text of a target language. The task are conventionally implemented as a cascaded system [1]: *(i)* the speech is first transcribed by Automatic Speech Recognition (ASR) model, *(ii)* the recognition results are fed to the Machine Translation (MT) model. In recent years, end-to-end methods [2, 3] attract considerable attention and achieve comparable or even better than its cascaded counterparts in certain applications [4]. Regardless of the approaches, ST systems are increasingly benefiting from the emergence of massively pretrained acoustic models [5] and language models [6].

A typical way of utilizing pretrained models is by fine-tuning the general-purposed model on task-specific training data. However, fine-tuning rewrites the parameters of pretrained model to favor a specific task at the risk of over-fitting and of forgetting useful generation skills, especially with limited training samples [7, 8]. One disadvantage of fine-tuning a multi-task model is that it adapts the model in a non-revertible manner so that the new model can no longer be shared across multiple tasks. Lightweight tuning methods such as adapter-tuning [9] or prefix-tuning [10] are proposed to efficiently adapt the pretrained models by only updating 'plug-in' modules inserted into the original model while freezing the rest. Such technique significantly reduces the amount of trainable parameters and preserve the generality of pretrained model. Existing study has shown that the lightweight-tuning technique could also potentially address problems such as forgetting or over-fitting.

In this paper, we investigate the effectiveness of prefix-tuning on the task of speech translation and propose a cross-modal prefix-tuning method leveraging acoustic feature as a supplementary channel to the MT module in a cascaded system. We freeze the parameters of both ASR and MT model to preserve their generality. Although our work is largely inspired by the prefix-tuning method [10] proposed for conditional text generation and the vision-based prefix [11], to our knowledge, we are the first to apply prefix-tuning technique to speech translation task. Our contributions are two folds: *(i)* to demonstrate the effectiveness and parameter-efficiency of cross-modal prefix-tuning on speech translation task for medium/low resource language pair, *(ii)* to demonstrate the robustness of cross-modal prefix-tuning in ASR outputs, which can mitigate the compounding errors between ASR and MT models.

2. RELATED WORK

Prompt-based approaches have been proposed as a lightweight tool to stimulate the knowledge in pretrained language models [12]. A prompt refers to a sequence of tokens, which values can be discrete or continuous, to be combined with the input sentence. Prompts can be designed to solve various tasks, presumed that those tasks could be predicted by the general language model, for example, sentiment classification. Prefix-tuning [10] prepends learnable task embeddings to the output of each layer in a pretrained model. Differing from prompts, which consist of discrete tokens, prefix-tuning employs continuous vectors that are not parameterized by the original model and allow more flexibility for learning the adaptation. The number of trainable parameters can be fur-

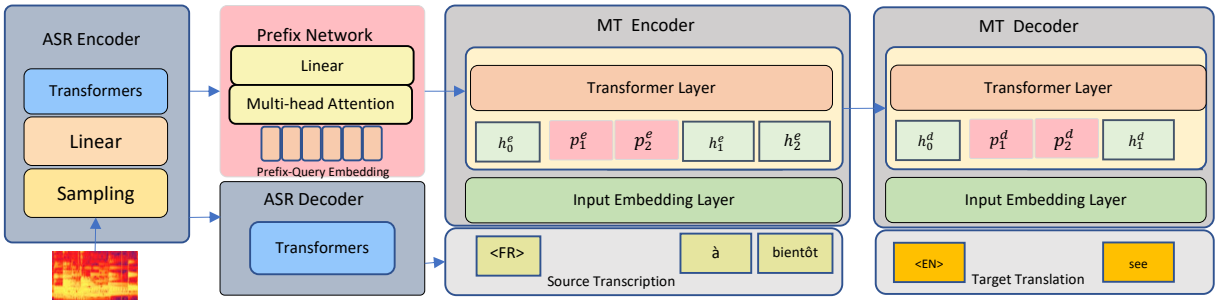


Fig. 1: Overview of the proposed cross-modal prefix-tuning for speech to text translation. Trainable components are highlighted in pink. **ASR Encoder** and **Decoder** are components of pretrained ASR model. **MT Encoder** and **Decoder** are components of pretrained mBART model. **Prefix Network** to generate prefixes from the contextual acoustic features

ther reduced by only prepending prefix to the input layer [13]. Extended from text-based methods, a cross-model prefix [11] is proposed for image-based question answering task. The prefix embeddings are generated based on the outputs of the vision encoder, which depends on the input images; thus these embeddings can be considered as sample-based prefixes.

Adapter [9, 14] is another popular approach to lightweight tuning of a pretrained model. Since the parameters of backbone model have been kept intact, the adapted model is less vulnerable to forgetting or over-fitting issues [7]. The speech community has also seen benefits from adapting pretrained language models using different tuning approaches. The encoder and decoder of an end-to-end speech translation model can be initialized with pretrained sequence-to-sequence models [6] with the help of fine-tuning [15] or adapter-tuning [16].

The other research line relates to building 'soft-connect' [17] between the modules of a cascaded speech translation model. The translation module could receive additional information supplementary to the 1-best ASR transcripts. Moreover, such soft-connect also allows errors to back-propagate to the ASR modules. One way to build such connect is by passing the posterior of predicted labels to the MT model's input layer [17], or by stacking the encoders of translation model and that of the ASR model [15] or by both [18].

3. CROSS-MODAL PREFIX

The proposed architecture is illustrated in Figure 1, which comprises of an ASR model, a multilingual MT model, and a Cross-modal Prefix Network. h^e_*, p^e_*, h^d_* and p^d_* stand for the hidden vectors [1] and prefix vectors in the encoder and decoder respectively.

Acoustic features (log filter-bank) are encoded by the acoustic encoder and then fed into ASR decoder to generate the speech transcripts of the source language. At the same time, the encoded acoustic features are passed through a Prefix Network to yield fixed-length prefixes that are later inserted to the output of each layer of the translation model. The Prefix Network is designed to be efficient, consisting of only sub-sampling and fully connected layer followed by multi-head attention.

Let X be the input audio features. The cross-modal prefix p_l at layer l is computed as:

$$p_{l,1\cdots N} = f^l_P(f_E(X), Q_l)$$

where f^l_P and f_E are respectively the function of Prefix Network and ASR encoder; Q_l is the $N \times D$ learnable prefix-query embeddings at layer l, with N being the prefix length, and D being the embedding dimension.

The hidden outputs of a given layer l for a sequence of length K, $H^l = [h^l_0, h^l_1, \cdots, h^l_K]$ is then modified by the corresponding prefix p_l as

$$\hat{H}^l = [h^l_0, p^l_1, p^l_2, \cdots, p^l_N, h^l_1, \cdots h^l_K]$$

Note that, we illustrate our study based on a multilingual MT model, which is implemented as a standard sequence-to-sequence architecture and distinguish language pairs based on specific language identifier in both encoder and decoder. Note that, h^l_0 refers to the language/task ID embedding, but it may vary with a different choice of pretrained model. During training, we only update parameters of prefix network while keeping the rest unchanged. It is of notable mention that this architecture is more similar to a cascaded pipeline as we still generate ASR transcriptions as intermediate outputs.

4. EXPERIMENT

4.1. Experiment Setting

We compare different adaptation techniques on ES-X and FR-X translation tasks using the recently published Mutli-lingual

[1] For simplicity, we ignore $*^e$ and $*^d$ when illustrating the construction of hidden outputs with prefixes.

	Training Set Size	Word Error Rate
ES	189 hours	15.3
FR	189 hours	17.1

Table 1: Average word error rate for ASR models of Spanish and French.

TEDx (MTEDx) Data Set [19] as the benchmark. For ES-X, there are four target languages: English (69 hours), French (11 hours), Portuguese (42 hours), and Italian(11 hours). For FR-X, there are three target languages: English (50 hours), Spanish (38 hours), Portuguese (25 hours). Among all directions, we identify ES-FR and ES-IT which have only 11 hours of training data as low-resource, while the other directions are considered as medium-resource. We use the checkpoint of mBART [2] that has been trained with many-to-many translation task. The English and French ASR models are trained using mono-lingual training data from MTEDx, after removing the duplicate subset of training data that overlaps among language pairs. We report the average word error rate in Table 1. The models have been implemented using ESPNet Toolkit [3] with standard configuration: 12 transformer layers for the encoder and 6 transformer layers for the decoder.

4.2. Baselines

- **Cascaded Model**: ASR and MT model (mBART) in a pipeline.
- **Cascaded Fine-tuned Model**: ASR and MT fine-tuned (100% of parameters) on speech-translation data.
- **E2E**: ESPNet implementation of transformer trained with multi-task objectives;
- **LNA** [15]: mBART initialized end-to-end model with only fine-tuning on parameters of LayerNorm and multi-head attention layers.
- **Adapter** [16]: Adapter-tuning with trainable adapter layers inserted in-between each transformer layer of pretrained mBART Model;
- **Adapter+** [16]: Added fine-tuning of the cross-attention module of mBART Decoder;
- **Prefix-tuning**: Trainable prefix added to both encoder and decoder of mBART;
- **CPT-Embedding Only**: Trainable prefix generated based on embeddings of acoustic feature. Prefixes are prepended to only the input layer of mBART;
- **CPT-Full**: Prefix being added to the outputs of all layers of mBART.

[2] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[3] https://github.com/espnet/espnet

4.3. Analysis

We summarize the results of ES-X and FR-X on the test set in Table 2. For prefix-tuning models, we apply free random seeds and run each setting in 10 runs. Reported performance is averaged over the multiple runs. We also include the number of trainable parameters averaged over all language pairs. Overall, prefix-tuning is able to achieve improvement over other tuning methods with significantly smaller sets of trainable parameters. On a closer look at the number of parameters, the capacity of mBART allows it to encode sufficient general knowledge to be shared across different languages, and thus shows a great potential in low-resource setting as long as the related knowledge could be encoded and leveraged. In comparison, end-to-end model simply fails in low-resource settings due to insufficient data for learning from scratch.

We observed that mBART (before fine-tuning) tends to be English-centric. Its performance is much stronger on X-to-En directions than other directions, even without any tuning. Although the model is pretrained in a many-to-many translation manner, English is still the dominating target language during training. As a result, the model exhibit strong English-centric behavior and fine-tuning on other language directions fails to show significantly improvement over the original model. We also observe in our experiment that mBART model (w/o fine-tuning) tends to generate English translation ignoring designated target language ID.

To investigate the behaviours of CPT vs fine-tuning, we conduct three experiments during testing – remove source or target language ID and randomly mask 20% tokens of ASR output. The results are shown in Table 3. It shows the fine-tuned mBART model is less sensitive to the removed source and target language ID, implying a possible overfitting to the designated task and domain. It is further verified by the fact that it becomes more vulnerable to random masks on the ASR outputs. In comparison, CPT relies more on the existing structural indicator (i.e., language IDs) to defend its specialization on given task. It is more robust to noises from the ASR outputs as the prefix embeddings are generated based on the acoustic features from the acoustic encoder.

4.4. Ablation Study

To understand the function of cross-modal prefix, we conduct ablation study for ES-EN, ES-FR, and FR-EN. It shows that for ES-EN and FR-EN, inserting prefix to the input layer seems to be sufficient. Encoder and decoder prefixes are contributes rather equally to the system's performance. In comparison, ES-FR prefers learnable prefix attached to each layer and relies more on the decoder. Considering that the embedding-only prefix relies on the transformer layer of pretrained model to yield prefixes for layers other than the input, it shows that the English-related attention of pretrained model is more adaptive as compared to its French counterpart.

	Trainable Parameter (M)	es-en	es-fr	es-it	es-pt	fr-es	fr-en	fr-pt	avg.
Cascade	0.0	17.52	8.11	6.04	7.61	12.67	23.39	5.53	8.85
Cascade (Fine-tuned)	610	20.95	14.46	16.12	23.19	24.07	23.83	25.12	17.66
E2E (ESPNET)	46	15.01	1.84	2.87	20.10	10.39	10.98	2.99	9.16
LNA[15]	104	18.38	9.67	8.46	19.62	21.37	21.91	22.03	17.34
Adapter[16]	6.3	15.82	13.83	17.12	23.40	26.14	26.17	25.07	21.07
Adapter+[16]	56.7	14.57	14.82	15.77	23.81	23.71	23.60	23.39	19.95
Prefix-tuning[10]	0.9	21.21	18.77	17.50	19.48	28.78	**32.61**	26.51	23.79
CPT-Embedding Only	0.5	23.85	18.07	18.97	23.83	27.63	31.29	27.19	24.41
CPT-Full	1.1	**23.88**	**20.03**	**19.62**	**26.08**	**29.13**	31.22	**29.64**	**25.28**

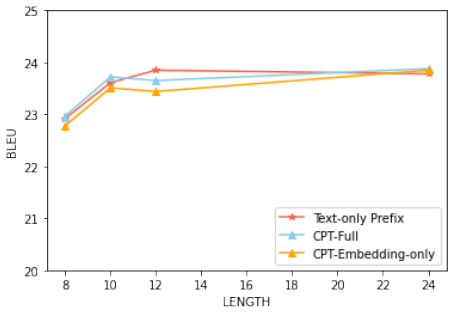
Table 2: ES-X and FR-X speech translation results on the test set of MTEDx in BLEU.

	Action	es-en	es-fr
CPT	Src lang. ID removal	3.95	6.17
	Tgt lang. ID removal	2.07	4.95
	20% Random Masking	6.23	7.40
Finetuning	Src lang. ID removal	-1.70	0.43
	Tgt lang. ID removal	1.38	2.01
	20% Random Masking	9.13	13.49

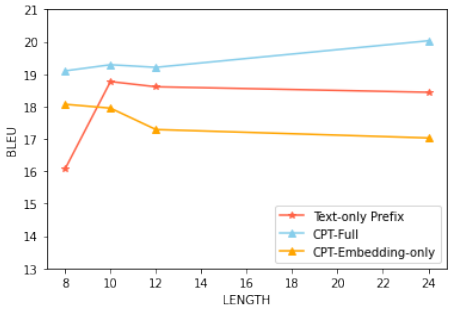
Table 3: Performance-drop (%) of CPT-Full and fine-tuned cascaded model when removing language IDs and introducing random noises for ES-FR and ES-EN translation.

	es-en	es-fr	fr-en
Full	23.88	20.03	31.22
-Layerwise	23.85	18.07	31.29
-Encoder	21.41	17.78	25.61
-Decoder	21.90	16.71	29.78

Table 4: Ablation Study of CPT-Full on ES-FR and ES-EN.



(a) ES-EN



(b) ES-FR

Fig. 2: The BLEU scores of prefix-tuning methods on ES-FR and ES-EN with varying lengths.

We then examine the impacts of prefix length. In Figure 2, we compare the BLEU score achieved by three prefix variants (i.e., Prefix-tuning, CPT-Full and CPT-Embedding) with varying prefix length on ES-FR and ES-EN translation tasks. It shows that, for ES-EN translation, all the three methods seems to reach their limit by using a length of 12. Simply increasing the prefix length only have minor influence on the performance. On low-resource setting (i.e., ES-FR), the influence of length is more significant. Embedding-only performs consistently poorer than the other two regardless of the choice of length. It is likely that layerwise prefix has more control over the acoustic information being passed from layers to layers. In comparison, CPF-Full demonstrated the potential of further improvement given larger capacity for the acoustic information. On the other hand, it shows that vanilla prefix-tuning is less sensitive to lengths.

5. CONCLUSION

We investigated cross-modal prefix-tuning for adapting multilingual translation model to bilingual speech translation task. The proposed prefix depends on both the given task and input speech. By providing layers of the translation model with direct access to the acoustic features through the prefixes, we show that prefix can effectively stimulate the knowledge of the pretrained model and achieve significant improvement on both medium- and low-resource settings.

6. REFERENCES

[1] Hermann Ney, "Speech translation: coupling of recognition and translation," in *1999 ICASSP*, 1999, pp. 517–520.

[2] Ye Jia, M. Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," *Proceedings of 2019 ICASSP*, pp. 7180–7184, 2019.

[3] S. Indurthi, HJ Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim, "End-end speech-to-text translation with modality agnostic meta-learning," in *Proceedings of 2020 ICASSP*, 2020, pp. 7904–7908.

[4] Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," in *Proceedings of the 2021 ACL-IJCNLP*, 2021.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[6] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *TACL*, vol. 8, pp. 726–742, 2020.

[7] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si, "On the effectiveness of adapter-based tuning for pretrained language model adaptation," in *Proceedings of the 2021 ACL-IJCNLP*, Online, Aug. 2021, pp. 2208–2222.

[8] Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng, "Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models," in *Proceedings of the 2021 EACL*, 2021, pp. 1121–1133.

[9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*. PMLR, 2019, pp. 2790–2799.

[10] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 2021 ACL-IJCNLP*, Online, Aug. 2021, pp. 4582–4597.

[11] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill, "Multimodal few-shot learning with frozen language models," *arXiv preprint arXiv:2106.13884*, 2021.

[12] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller, "Language models as knowledge bases?," in *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, 2019, pp. 2463–2473.

[13] Brian Lester, Rami Al-Rfou, and Noah Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[14] Wenjuan Han, Bo Pang, and Ying Nian Wu, "Robust transfer learning with pretrained language models through adapters," in *Proceedings of the 2021 ACL-IJCNLP*, Online, Aug. 2021, pp. 854–861.

[15] Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "Multilingual speech translation from efficient finetuning of pretrained models," in *Proceedings of the 2021 ACL-IJCNLP*, Online, Aug. 2021, pp. 827–838.

[16] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier, "Lightweight adapter tuning for multilingual speech translation," in *Proceedings of the 2021 ACL-IJCNLP)*, Aug. 2021, pp. 817–824.

[17] Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong, "Bridging the modality gap for speech-to-text translation," *arXiv preprint arXiv:2010.14920*, 2020.

[18] Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu, "Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders," in *Proceedings of 2021 ACL-IJCNLP*, Online, Aug. 2021, pp. 2619–2630.

[19] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post, "The Multilingual TEDx Corpus for Speech Recognition and Translation," in *Proc. Interspeech 2021*, 2021, pp. 3655–3659.