Modality Adaption or Regularization? A Case Study on End-to-End Speech Translation

Yuchen Han[1], Chen Xu[1], Tong Xiao[1,2]*, Jingbo Zhu[1,2]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
hanyuchen114@gmail.com,xuchennlp@outlook.com
{xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract

Pre-training and fine-tuning is a paradigm for alleviating the data scarcity problem in end-to-end speech translation (E2E ST). The commonplace "modality gap" between speech and text data often leads to inconsistent inputs between pre-training and fine-tuning. However, we observe that this gap occurs in the early stages of fine-tuning, but does not have a major impact on the final performance. On the other hand, we find that there has another gap, which we call the "capacity gap": high resource tasks (such as ASR and MT) always require a large model to fit, when the model is reused for a low resource task (E2E ST), it will get a sub-optimal performance due to the overfitting. In a case study, we find that the regularization plays a more important role than the well-designed modality adaption method, which achieves 29.0 for en-de and 40.3 for en-fr on the MuST-C dataset. Code and models are available at https://github.com/hannlp/TAB.

1 Introduction

End-to-end speech translation (E2E ST) employs a direct model to translate source language speech into target language text, which has low latency and can avoid the "error propagation" problem in traditional cascade methods (Weiss et al., 2017). However, compared to automatic speech recognition (ASR) and machine translation (MT) models used in cascade methods, E2E ST models typically have limited training data (Cattoni et al., 2021), which can result in sub-optimal performance.

Transferring the knowledge from the related tasks (e.g. ASR and MT) is a widely-used approach for E2E ST to achieve optimal performance (Tang et al., 2021; Zhang et al., 2022a). However, the difference between tasks and data makes the transfer process more challenging (Wang et al., 2020). The inconsistency of length and representation between speech and text leads to the "modality gap" (Liu et al., 2020), which exists in scenarios where the inputs of the model change, such as in the pre-training and fine-tuning (PT-FT) paradigm (Xu et al., 2021) or in the multi-task learning (MTL) methods (Ye et al., 2021). Thus, the connectionist temporal classification (CTC) (Graves et al., 2006) based adapters (Liu et al., 2020; Xu et al., 2021) have been proposed to transform the original speech output into a text-like sequence. Recently, consistency training methods have achieved promising results by using a better branch, such as the mix-up branch (Fang et al., 2022) or the text branch (Cheng et al., 2022), to promote cross-modal learning and support the original speech output. However, we find that the "modality gap" does not exist throughout the training process in Figure 1. While consistency training methods have some regularization function (Liang et al., 2021; Guo et al., 2022) to help the model overcome over-fitting and be fully trained. A natural question arises: *Are modality adaption methods always effective when an E2E ST model is fully trained?*

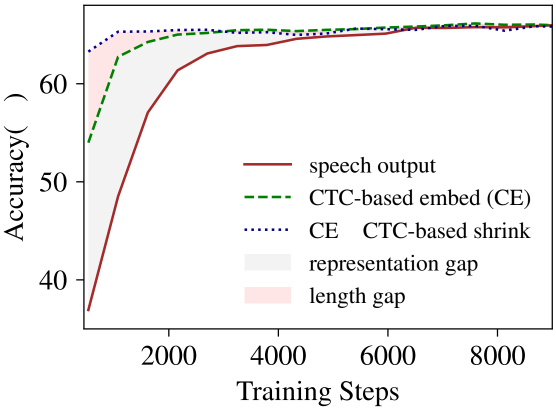

Figure 1: The modality gap in the PT-FT paradigm. We decoupled the encoder to speech encoder and text encoder (Wang et al., 2020; Xu et al., 2021), and compared the accuracy curve of fine-tuning for different inputs of the text encoder. We find that the modality gap only exists at the initial stage of the fine-tuning.

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1340–1348
July 9-14, 2023 ©2023 Association for Computational Linguistics

In this work, we aim to investigate how much of the improvement is due to the modality adaption or regularization methods. To achieve this, we adopt the PT-FT and encoder decouple paradigm and establish a framework that incorporates adjustable modality adaption and consistency training methods. Through extensive experiments on the MuST-C en-de and en-fr benchmarks, we observe that:

- The modality adaption method in PT-FT only accelerates the early phase of fine-tuning, but does not provide a significant improvement for a fully trained model.
- We obtained 29.0 and 40.3 on the MuST-C en-de and en-fr datasets, but regularization played a major role, which confirming that the "capacity gap" is more severe than the "modality-gap" in E2E ST.

2 Our Case Study: TAB

2.1 Architecture

The E2E ST corpus $\mathcal{D}_{ST}$ usually consists of three parts: the source language speech $\boldsymbol{s}$, the corresponding transcripts $\boldsymbol{x}$, and the target translation $\boldsymbol{y}$. The overall framework, as shown in Figure 2, consists of a speech encoder and a shared transformer.

Speech Encoder. The speech encoder encodes the source language speech $\boldsymbol{s}$ into speech output $\boldsymbol{h}$:

$$\boldsymbol{h} = \mathrm{ENC}_{\text{speech}}(\boldsymbol{s}; \theta_s) \tag{1}$$

We employ the CTC loss at the top of the speech encoder (Wang et al., 2020; Liu et al., 2020; Xu et al., 2021) to predict an alignment path $\boldsymbol{\pi}$ by speech output $\boldsymbol{h}$ based on a conditionally independent assumption $p(\boldsymbol{\pi}|\boldsymbol{h}) = \prod_t^{|\boldsymbol{h}|} p(\pi_t|\boldsymbol{h_t})$:

$$p(\pi_t|\boldsymbol{h_t}) = \mathrm{Softmax}(\mathrm{Linear}(\boldsymbol{h_t}; \theta_{ctc})) \tag{2}$$

where $\pi_t \in \mathcal{V}^+$, which is an extended source language vocabulary by introducing a "blank" token. The path $\boldsymbol{\pi}$ can be mapped to the transcript $\boldsymbol{x}$ by removing all repeated labels and blanks, such an operation is called β. The CTC loss is defined as the negative log probability of all possible alignment paths $\beta^{-1}(\boldsymbol{x})$ between $\boldsymbol{h}$ and $\boldsymbol{x}$:

$$\mathcal{L}_{CTC} = -\sum_{\boldsymbol{\pi} \in \beta^{-1}(\boldsymbol{x})} \mathrm{log} p(\boldsymbol{\pi}|\boldsymbol{h}) \tag{3}$$

Shared Transformer. The shared transformer accepts two inputs: the original branch $\boldsymbol{o}$ and the auxiliary branch $\boldsymbol{a}$, both of equal length. We leave the details of these to the Section 2.2. The encoder and decoder of the shared transformer are utilized to obtain the final predictions $\mathcal{P}_j = p(y_j|\boldsymbol{y_{<j}}, \boldsymbol{o}; \theta_t)$ and $\mathcal{Q}_j = p(y_j|\boldsymbol{y_{<j}}, \boldsymbol{a}; \theta_t)$, respectively. The cross-entropy losses are then calculated as follows:

$$\mathcal{L}_{CE_o} = -\sum_{j=1}^{|y|} \mathrm{log}\mathcal{P}_j \tag{4}$$

$$\mathcal{L}_{CE_a} = -\sum_{j=1}^{|y|} \mathrm{log}\mathcal{Q}_j \tag{5}$$

2.2 Tuning with Auxiliary Branch (TAB)

The speech encoder and shared transformer are initialized with pre-trained ASR and MT models. During fine-tuning, we aim to build a text-like auxiliary branch which includes some textual representations like Fang et al. (2022); Zhang et al. (2022b) for modality adaption, and provide an adjustable probability to control the degree of it. To obtain the textual embedding, we utilize the ability of the CTC alignment, where $\pi_t = \mathrm{argmax}(p(\pi_t|\boldsymbol{h_t}))$ is an id of $\mathcal{V}^+$ that denotes the corresponding CTC-predicted token of the speech feature $\boldsymbol{h_t}$.

Shrink. To eliminate the effect of too many "blank" positions in the sequence, we first average the consecutively repeated features (e.g. $\pi_i = ... = \pi_j = c_{i \to j}$) in the speech output $\boldsymbol{h}$ to obtain the original branch $\boldsymbol{o}$, where $\boldsymbol{o_k} = (\boldsymbol{h_i} + ... + \boldsymbol{h_j}) \cdot \frac{1}{j-i}$.

Copy & Replace. We copy $\boldsymbol{o}$ to a new sequence $\boldsymbol{a}$ to ensure that the auxiliary branch has the same length as the original branch. Each position $\boldsymbol{a_k}$ in the new sequence is then replaced with its CTC predicted embedding $\mathrm{Embedding}(c_{i \to j})$ with a probability p^* if $c_{i \to j}$ is not a "blank". Here, $\mathrm{Embedding}$ is an embedding matrix initialized by the pre-trained MT model. The replacement probability p^* can be adjusted to control the degree of modality adaption, which can be a fixed or a dynamic value, as discussed in Section 4.3. It is important to note that the auxiliary branch provides a regularization function due to the replacement operation or dropout (Liang et al., 2021). This effect will be further discussed in Section 4.4.

Fine-tuning strategy. To utilize the auxiliary branch, a consistency loss is introduced to enforce consistency between two output distributions:

$$\mathcal{L}_{Cons} = \sum_{j=1}^{|y|} \mathcal{D}(\mathcal{P}_j, \mathcal{Q}_j) \tag{6}$$

where $\mathcal{D}$ denotes the loss term. The final loss used in TAB is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE_o} + \mathcal{L}_{CE_a} + \lambda \mathcal{L}_{CTC} + \alpha \mathcal{L}_{Cons} \tag{7}$$

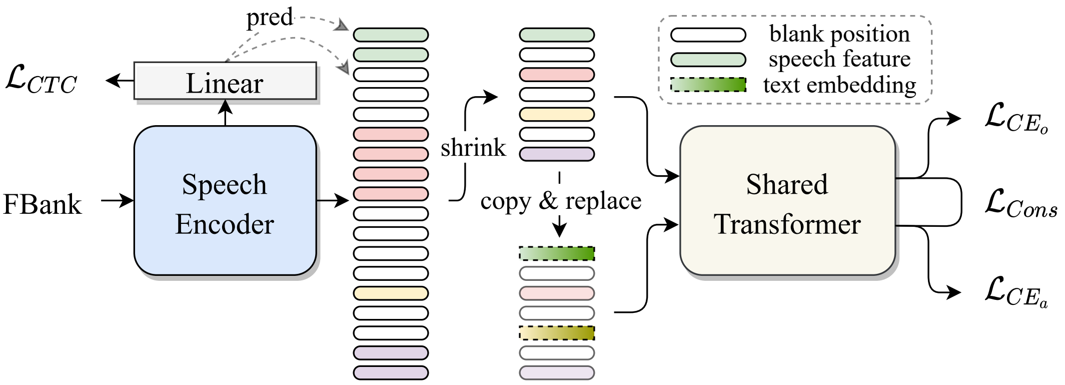

Figure 2: An illustration of TAB. The speech output is first shrunk based on the CTC predicted path, then copied and replaced by text embedding with a probability p^* to obtain the auxiliary branch. Two branches are fed into a shared transformer, and the final predictions are kept consistent by consistency loss.

3 Experimental Setup

Datasets and Pre-processing. We conducted our experiments on the MuST-C (Gulati et al., 2020) dataset for two language directions: En-De and En-Fr. The dev set was used for validation, and the tst-COMMON set was used for reporting our results. For training the English ASR model, we used the LibriSpeech (Panayotov et al., 2015) dataset. The WMT16 En-De and WMT14 En-Fr datasets were used to train the MT models. Table 1 presents the statistics of all the datasets used in our pre-training and fine-tuning processes.

Dataset	ASR(H)	MT(S)	ST(H/S)
En-De	960	4.5M	408/234K
En-Fr	960	36M	492/280K

Table 1: Statistics of ASR, MT and ST data for two language pairs, where H denotes hours and S denotes sentences.

We preprocessed the speech input by extracting 80-dimensional log-mel filterbank features and removing utterances with more than 3,000 frames. The vocabulary, which has a size of 10k, is shared between the source and target languages and was trained using the SentencePiece (Kudo and Richardson, 2018) model from the MuST-C dataset.

Model Configuration. All experiments were implemented using the fairseq toolkit (Ott et al., 2019). Two convolutional layers with a stride of 2 were introduced to downsample the input speech features. We used the Conformer (Gulati et al., 2020) as our speech encoder, which consists of 12 layers. Both the text encoder and decoder in the shared transformer have 6 layers. Each layer in our model has 512 hidden units, 2048 feed-forward size, and 8 attention heads. The ASR and MT models were pre-trained with external data and fine-tuned with the MuST-C dataset.

Training and Inference. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.997$ in MT, while $\beta_2 = 0.98$ in ASR and ST, respectively. During ASR pre-training, each batch has up to 800k frames, and the learning rate and warmup were 1.4e-3 and 10000. During MT pre-training, each batch has up to 33k tokens, and the learning rate and warmup were 1e-3 and 8000. During ST fine-tuning, each batch has up to 320k frames, and the learning rate and warmup were 7e-4 and 4000. The hyper-parameter λ was set to 0.3 for both pre-training and fine-tuning. We used dropout with a ratio of 0.1 during pre-training and 0.15 during fine-tuning, and label smoothing with a value of 0.1. All training was stopped early if the loss (ASR and MT) or BLEU (E2E ST) on the dev set did not improve for twenty epochs. During inference, we averaged the model parameters of the best 10 checkpoints for evaluation. We used a beam search with a beam size of 5 for all models. We reported the case-sensitive SacreBLEU (Post, 2018). All models were trained on 4 Titan RTX GPUs.

4 Results and Discussion

4.1 Which type of consistency loss is best?

The choice of an appropriate consistency loss is crucial in leveraging the knowledge from the auxiliary branch, whether it is due to modality adaption or regularization. We conducted experiments with different loss terms with $\alpha = 1$ and $p^* = 0.2$.

Loss term	BLEU		
	dev	tst-COMMON	avg.
None ($\alpha = 0$)	27.69	27.99	27.84
JSD	27.76	28.49	28.13
KL$_{orig \rightarrow aux}$	28.47	28.49	28.48
KL$_{aux \rightarrow orig}$	28.26	28.78	28.52
bi-KL	28.43	28.78	**28.61**

Table 2: BLEU scores on MuST-C en-de dev and tst-COMMON sets with different consistency loss terms.

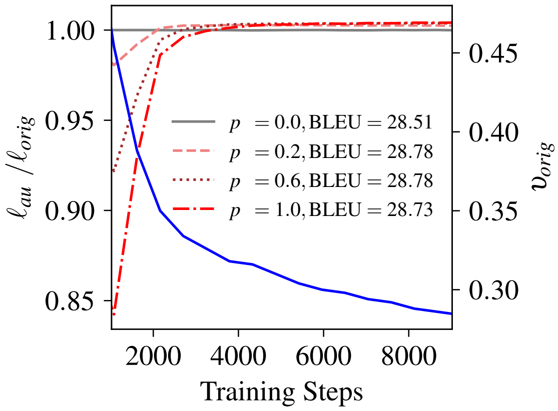

Figure 3: The ratios of the auxiliary branch loss to the original branch loss ℓ_{aux}/ℓ_{orig} at each training step under different fixed values of p^*. Additionally, we show the uncertainty of the original branch υ_{orig} (represented by the blue line) at the same time.

As shown in Table 2, the results indicate that a consistency loss is necessary to improve performance. The Jensen-Shannon Divergence (JSD) loss and the unidirectional-KL loss were found to be slightly worse than the bidirectional-KL (bi-KL) loss. Therefore, we selected the bi-KL loss for the subsequent experiments.

4.2 Whether & when the modality gap exist?

In Figure 3, we present the ℓ_{aux}/ℓ_{orig} curve of fine-tuning, which represents the ratio of the auxiliary branch loss ℓ_{aux} to the original branch loss ℓ_{orig}. The only difference between the two branches is the replaced textual embeddings for modality adaption in the auxiliary branch. We investigate the effect of this operation during fine-tuning under different replacement probabilities.

Our results show that, in the early fine-tuning phase, the ℓ_{aux} always lower than ℓ_{orig} when $p^* > 0$, indicating that the model expects some textual representations like the input in pre-training. However, in the later fine-tuning process, the ℓ_{aux} is slightly higher than the ℓ_{orig}, suggesting that the model starts to get confused about replacement operations due to the introduction of some kind of noise by the destruction of the original sequence. Moreover, we find that the maximum $p^* = 1.0$ always has the lowest ratio at the beginning and the highest ratio at other times. This confirms that the modality gap exists but not throughout the entire fine-tuning process.

4.3 Does the modality adaption help?

We experimented with different fixed replacement ratios p^* over the range $[0.0, 0.2, 0.6, 1.0]$ under $\alpha = 1.0$ in Figure 3. Our results showed that the method with modality adaption $p^* > 0$ consistently outperformed $p^* = 0$. Furthermore, as observed in Section 4.2, there exists a "tipping point" in the fine-tuning process where the modality gap disappears. Before this point, we should use a higher p^*, while a lower p^* should be more effective after this point, rather than using a fixed value. We discovered that the uncertainty of the original branch υ_{orig}, which is defined by the normalized entropy of P_j [1], is strongly related to this point, as shown in Figure 3:

$$\upsilon_{orig} = -\frac{1}{\log V} \cdot \frac{1}{|y|} \sum\nolimits_{j=1}^{|y|} \mathcal{P}_j \log \mathcal{P}_j \quad (8)$$

where V is the vocabulary size. We then proposed a dynamic replacement probability derived from υ_{orig} at each step: $p^* = \gamma \cdot \upsilon_{orig}$, where γ is a hyper-parameter set to 0.5 in all experiments. When we use a dynamic replacement ratio of $p^* = 0.5 \cdot \upsilon_{orig}$, we denote it as $p^* = \upsilon$. By adopting this dynamic replacement ratio, we achieved a BLEU score of 28.87 on the MuST-C en-de dataset.

4.4 Is modality adaption always effective?

The consistency loss has been demonstrated to have a regularization effect on two branches with different outputs (Liang et al., 2021; Guo et al., 2022). When $p^* = 0$, there is no modality adaption through the replacement operation, but the dropout still causes different outputs for the two branches, the TAB degenerates into a pure regularization method. The hyper-parameter α can be used to control the degree of regularization, where a higher α indicates stronger regularization. By increasing α to 5, it is observed that the gap between the modality adaption method ($p^* = 0.2$ or $p^* = \upsilon$) and the pure regularization method

[1] More precise definitions are $P_j(t)$ and $\upsilon_{orig}(t)$, with the symbol of the training step "t" omitted for brevity.

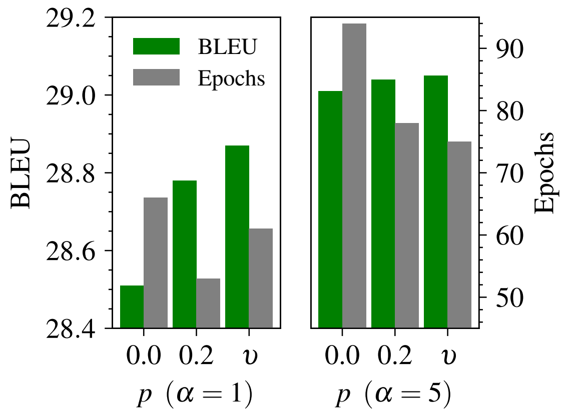

Figure 4: BLEU scores and the epochs required for convergence with different values of α and p^* on the MuST-C en-de dataset. The υ denotes the use of an uncertainty-based ($\gamma = 0.5$) dynamic value for p^*.

($p^* = 0$) decreases (29.05 vs 29.01) in Figure 4, although the pure regularization method always required more epochs to converge, which can be attributed to more complete training. These findings confirm that the modality adaption method on PT-FT can accelerate the early phase of fine-tuning, but does not significantly enhance the final performance when a sufficient level of regularization has been applied to the model. This also highlights the issue of over-fitting being a more serious problem in E2E ST than the "modality gap", and that better regularization and a longer time for fine-tuning can help to eliminate the "modality gap" problem in the PT-FT paradigm.

4.5 Final results and comparison of methods

The results for the MuST-C dataset are shown in Table 3. The modality adaption method shows an improvement of 1.0 and 0.7 BLEU points over our CTC baseline at a lower consistency level ($\alpha = 1$). However, the pure regularization method with $\alpha = 5$ slightly outperforms it (+0.1 BLEU), and outperforms all other methods designed for modality adaption ($+0.1 \sim 1.2$ BLEU), except those using HuBERT as the feature extractor. When we combined our modality adaption method with a higher level of consistency ($\alpha = 5$), further improvement can still be achieved, but not consistently in all languages. Our hypothesis is that the replacement operation in TAB not only releases the modality gap in the early fine-tuning phase but also introduces noise in the later stages. This noise can bring better regularization performance (Guo et al., 2022; Gao et al., 2022) when a higher consistency level is given. In general, regularization brings more improvement in TAB than modality adaption, and a better regularization is helpful for E2E ST models to be fully trained.

Methods	BLEU	
	En-De	En-Fr
SATE (Xu et al., 2021)	28.1	-
STEMM† (Fang et al., 2022)	28.7	37.4
ConST† (Ye et al., 2022)	28.3	38.3
M^{3}ST†† (Cheng et al., 2022)	29.3	38.5
WACO† (Ouyang et al., 2022)	28.1	38.1
AdaTrans (Zeng et al., 2022)	28.7	38.7
CRESS† (Fang and Feng, 2023)	28.9	-
CRESS†† (Fang and Feng, 2023)	**29.4**	40.1
baseline (CTC)	27.9	39.1
TAB ($p^* = \upsilon, \alpha = 1$)	28.9	39.8
TAB ($p^* = 0, \alpha = 5$)	**29.0**	39.9
TAB ($p^* = \upsilon, \alpha = 5$)	**29.0**	**40.3**

Table 3: BLEU scores on the MuST-C tst-COMMON set. Methods marked with † employ self-supervised pre-training models that were trained on unsupervised speech data instead of ASR data. † refers to Wav2Vec2.0 and †† refers to HuBERT.

5 Conclusion

Through a case study, we have demonstrated that the "modality gap" in the PT-FT paradigm for E2E ST is only present in the early stages of fine-tuning. Although a modality adaption method can accelerate the convergence speed, it does not significantly improve the final performance of a fully trained model. However, the over-fitting and "capacity gap" are more critical issues in E2E ST, and a good regularization technique can help in fully training the model.

Acknowledgement

The authors would like to thank anonymous reviewers for their insightful comments. This work was supported in part by the National Science Foundation of China (No. 62276056), the National Key R&D Program of China, the China HTRD Center Project (No. 2020AAA0107904), the Natural Science Foundation of Liaoning Province of China (2022-KF-16-01), the Yunnan Provincial Major Science and Technology Special Plan Projects (No. 202103AA080015), the Fundamental Research Funds for the Central Universities (Nos. N2216016, N2216001, and N2216002), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No. B16009).

Limitations

Although our work has achieved high performance, there are still some limitations that need to be addressed in future work:

- Our work was only carried out under the assumption that there is sufficient ASR and MT data, and relied on transcriptions for CTC loss to perform alignment and predict. This assumption may not hold in real-world scenarios, and we plan to investigate the performance of our approach under more diverse data conditions in the future.

- We only attempted to feed two branches into the shared transformer in order to ensure a fair comparison between the pure regularization method and the modality adaption method. However, this approach may have resulted in sub-optimal regularization performance compared to methods that feed all branches into the whole model, as demonstrated by Liang et al. (2021); Guo et al. (2022); Gao et al. (2022).

References

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Comput. Speech Lang.*, 66:101155.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2022. M3ST: mix at three levels for speech translation. *CoRR*, abs/2212.03657.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. *CoRR*, abs/2305.08706.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7050–7062. Association for Computational Linguistics.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-simcut: A simple strategy for boosting neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3938–3948. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

Dengji Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2022. Prediction difference regularization against perturbation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7665–7675. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10890–10905.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. Bridging the modality gap for speech-to-text translation. *CoRR*, abs/2010.14920.

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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. WACO: word-aligned contrastive learning for speech translation. *CoRR*, abs/2212.09359.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Yun Tang, Juan Miguel 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4252–4261. Association for Computational Linguistics.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9161–9168. AAAI Press.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2625–2629. ISCA.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2619–2630. Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2267–2271. ISCA.

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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5099–5113. Association for Computational Linguistics.

Xingshan Zeng, Liangyou Li, and Qun Liu. 2022. Adatrans: Adapting with boundary-based shrinking for end-to-end speech translation. *CoRR*, abs/2212.08911.

Yuhao Zhang, Chen Xu, Bojie Hu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2022a. Improving end-to-end speech translation by leveraging auxiliary speech and text data. *CoRR*, abs/2212.01778.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1663–1676. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.