Mutual-Learning Improves End-to-End Speech Translation

Jiawei Zhao[1]*, **Wei Luo**[3], **Boxing Chen**[3], **Andrew Gilman**[1,2]
[1]School of Natural and Computational Sciences, Massey University, New Zealand
[2]PlantTech Research Institute, New Zealand
[3]Alibaba Inc.
j.zhao2@massey.ac.nz
andrew@pri.co.nz
{muzhuo.lw, boxing.cbx}@alibaba-inc.com

Abstract

A currently popular research area in end-to-end speech translation is the use of knowledge distillation from a machine translation (MT) task to improve the speech translation (ST) task. However, such scenario obviously allows only a one way transfer, limiting the overall effectiveness of the approach by the performance of the pre-trained teacher model. Therefore, we pose that in this respect knowledge distillation-based approaches are sub-optimal. We propose an alternative—a trainable mutual-learning scenario, where the MT and ST models are collaboratively trained and are considered as peers, rather than teacher/student. This allows us to improve the performance of end-to-end ST more effectively than with a teacher-student paradigm. As a side benefit, performance of the MT model also improves. Experimental results show that in our mutual-learning scenario, models can effectively utilise the auxiliary information from peer models and achieve compelling results on MuST-C datasets.

1 Introduction

Speech translation (ST) aims to translate speech signals into a foreign language. It is a multi-modality task, closely related to automatic speech recognition (ASR) and machine translation (MT). ST has a wide range of applications, such as video subtitling (Saboo and Baumann, 2019), real-time lecture translation (Müller et al., 2016), and protection of endangered languages (Bansal et al., 2017).

Despite the recent success in end-to-end (E2E) models, currently such systems still face the issue of labelled training data insufficiency (Sperber and Paulik, 2020). A recent popular advance in E2E ST is the use of knowledge distillation (KD), which can provide an effective paradigm for transferring knowledge from rich-resource to low-resource tasks (Liu et al., 2019; Gaido et al., 2020). Under such paradigm the MT model is considered a teacher that guides the ST model, considered a student learning from the teacher. We pose that one-way knowledge transfer in a strict teacher-student relationship maybe sub-optimal for the following reasons: 1. Since the MT model is frozen in this one-way knowledge transfer scenario, the success of knowledge transfer and hence the performance of the ST task is constrained by the performance of the pre-trained MT model; 2. There is a modality gap between speech and text inputs of the two models, with speech input also containing inherent speaker variability.

Motivated to address the issues mentioned above, we set out to improve ST and MT tasks by training them jointly. Instead of freezing teacher model, we introduce a mutual-learning paradigm, which regards ST and MT models as peers that learn collaboratively, aiming to iteratively learn and share the knowledge between the two models. Originally, mutual-learning was proposed to leverage information from multiple models and allow effective dual knowledge transfer in image processing tasks (Zhang et al., 2018; Zhao et al., 2021). We leverage this idea and adapt it to sequence tasks. Our main contributions are: 1. We propose a jointly-trainable mutual-learning paradigm, which improves the distillation method by training together. The search space of MT and ST are both enlarged, providing the potential for a more robust local optima. 2. We further improve our mutual-learning method by integrating cyclical annealing schedule, which alleviates the KL vanishing problem suffered by many time-series tasks (Fu et al., 2019; Bowman et al., 2016; Higgins et al., 2017). 3. We implement extensive experiments on MuST-C En-Fr, En-Es datasets and illustrate the advantage of our model by empirically comparing with a cascaded model, a knowledge distillation (KD) model and a multi-task learning (MTL) model. The ex-

* Majority of this work was conducted during Jiawei Zhao's research internship at DAMO Academy, Alibaba.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3989–3994
November 7–11, 2021.

perimental results show our model can effectively leverage the transcript and the auxiliary MT task, and we provide competitive results in all experiments. In addition, as a side benefit, the performance of the MT model also improves.

2 Model Description

2.1 End-to-End Speech Translation

E2E ST learns a single model, which directly maps features extracted from speech signal to a target language text sequence (Duong et al., 2016; Weiss et al., 2017). More concretely, given a sample pair (x, y) from the training set D corresponding to speech signal features and translated target sentence, the ST model is trained by minimising the negative log likelihood (NLL) loss, L:

$$L = -\sum_{(x,y)\in D} logP(y|x;\theta) \quad (1)$$

E2E models consist of an encoder that encodes speech input as an intermediate representation, and a decoder that decodes this intermediate representation to a probability distribution over the target text feature space. In the past, the encoder and decoder were based on recurrent neural network architecture, but most recent work utilises Transformer-based architecture (Berard et al., 2016; Weiss et al., 2017; Di Gangi et al., 2019b; Zhang et al., 2019; Vila et al., 2018).

2.2 Mutual-Learning

Model definition: Given a parallel data sample (x_i, s_i, y_i) from input speech features X, input language text features S and target text features Y, and an ST model M_{st} and an MT model M_{mt}, the output probabilities are given by:

$$p_{st} = M_{st}(x_i) \quad (2)$$

$$p_{mt} = M_{mt}(s_i) \quad (3)$$

Our training loss has two components: a traditional supervised reconstruction loss and a mimicry loss that aligns the output posterior distributions between the two models. We adopt Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) as the mimicry loss, aiming to reduce the distance of outputs of ST and MT systems, effectively encouraging them to mimic each other. Since KL divergence is asymmetric, we include it calculated in both directions:

$$KL_1 = KL(p_{mt}||p_{st}) = \sum_{j=1}^{N} p_{mt}^j \ln \frac{p_{mt}^j}{p_{st}^j} \quad (4)$$

$$KL_2 = KL(p_{st}||p_{mt}) = \sum_{j=1}^{N} p_{st}^j \ln \frac{p_{st}^j}{p_{mt}^j} \quad (5)$$

where N represents the length of the output sentence. We adopt NLL loss as the reconstruction loss, denoted by LC_{st} for ST and LC_{mt} for MT:

$$LC_{st} = -\sum_{i=1}^{N} y_i \ln (p_{st}^i|y_i) \quad (6)$$

$$LC_{mt} = -\sum_{i=1}^{N} y_i \ln (p_{mt}^i|y_i) \quad (7)$$

where y_i denotes the i_{th} token in the target sentence. The overall mutual-learning training loss is a combination of the weighted mimicry loss and the reconstruction losses, as described by Eq. 8. The proposed mutual-learning scenario is illustrated in Figure 1.

$$L_{ml} = \beta(KL_1 + KL_2) + LC_{st} + LC_{mt} \quad (8)$$

2.3 Training Strategy

Algorithm 1 Training Strategy

Input: training set, ST network parameters θ_{st} (with ASR pre-trained encoder), pre-trained MT network parameters θ_{mt}

repeat

$t = t + 1$

1. Compute p_{st} and p_{mt} for one mini batch
2. Freeze θ_{mt}, compute the gradient and update θ_{st}

$$\theta_{st} \longleftarrow \theta_{st} + lr * \frac{\partial L_{ml}}{\partial \theta_{st}} \quad (9)$$

3. Upate p_{st} and p_{mt}
4. Freeze θ_{st}, compute the gradient and update θ_{mt}

$$\theta_{mt} \longleftarrow \theta_{mt} + lr * \frac{\partial L_{ml}}{\partial \theta_{mt}} \quad (10)$$

until convergence

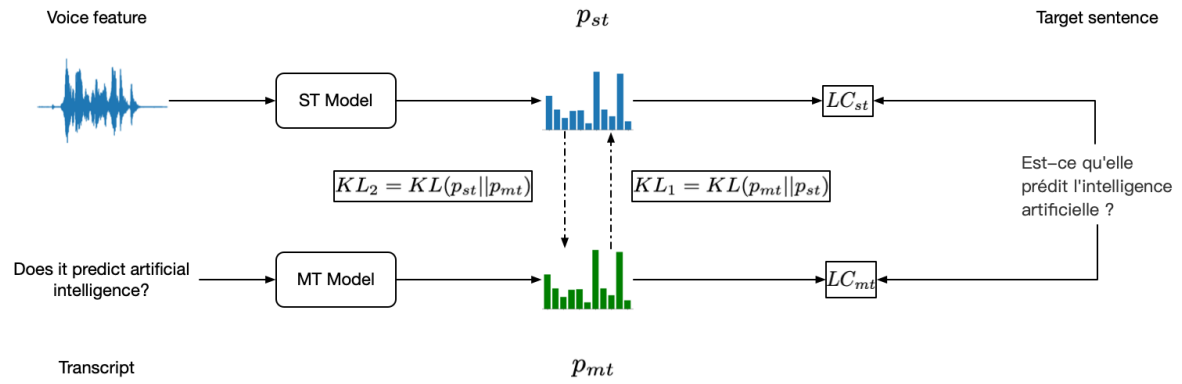

Figure 1: Illustration of the proposed deep mutual-learning paradigm. The training objective contains four separate components, the reconstruction losses of ST and MT (LC_{st} and LC_{mt}) and KL divergence between outputs of ST and MT (KL_1 and KL_2).

The training process is described by Algorithm 1. We propose to train ST and MT models iteratively until convergence. In each iteration, there are two steps: 1. MT model is frozen and the parameters of ST model are updated; 2. ST model is frozen and the parameters of MT model are updated.

KL vanishing issue: Leveraging KL divergence for mimicry loss in our mutual-learning strategy can suffer from the vanishing issue, which has been observed in other applications, for example in variational auto-encoders (Fu et al., 2019). We mitigate this by adopting a cyclical annealing schedule for β, which has been proposed for this purpose in the context of variational auto-encoders (Fu et al., 2019). More concretely, β in Eq. 8 changes periodically during training iterations, as described by Eq. 11:

$$\beta_t = \begin{cases} \frac{r}{RC}, & r <= RC \\ 1, & r > RC \end{cases} \quad (11)$$

where t represents the current training iteration and r is defined as:

$$r = mod(t-1, C) \quad (12)$$

The training process is effectively split into many cycles with each cycle lasting C iterations. In each cycle β_t progressively increases from 0 to 1 during RC iterations and then stays at 1 for the remaining $(1-R)C$ iterations. With $R = 0.5$ and $C = 5000$, we are able to mitigate KL vanishing issue and train.

3 Experiments

3.1 Dataset

We evaluate the proposed framework on the popular MuST-C multilingual speech translation corpus[1] (Di Gangi et al., 2019a), using the two most-used language pairs: English-to-French (En-Fr) and English-to-Spanish (En-Es). En-Fr dataset contains 500 hours of speech and 280k sentences. En-Es dataset contains 504 hours of speech and 270k sentences.

Pre-processing We implement the same data pre-processing steps as described in *fairseq* speech-to-text framework (Wang et al., 2020). Specifically, we extract 80-channel log Mel-filterbank features. The training samples that are larger than 3000 frames are removed. For both, input and target texts, we employ newly proposed subword regularisation method (Kudo, 2018) to build a vocabulary with a size of 8000. We also experiment with a jointly-trained shared vocabulary of size 8000.

3.2 Architecture and Evaluation Details

For ST task we use a stack of 2 1D convolutional layers (kernel size 5, stride 2), followed by 12 Transformer layers of size 2048 as the encoder. We use 6 stacked Transformer layers with size 512 as the decoder. For MT task we use 12 stacked Transformer layers with size 2048 as the encoder and 6 stacked transformer layer with size 2048 as the decoder. Evaluation is based on the standard implementation of BLEU score, SACREBLEU (Post, 2018), with beam size of 5. The maximum number of tokens in each batch is set to 40000.

[1] https://ict.fbk.eu/must-c/

4 Results and Analysis

4.1 Comparison with a Cascaded Model

To form a cascaded model, we first train a Transformer-based E2E ASR model using speech inputs and English transcripts. We then train an MT model using English transcripts and target sentences. In inference mode, we first use ASR to generate intermediate text representation, then we pass this to the MT system and calculate the output probabilities on the target language vocabulary.

As shown in Table 1, our mutual-learning-based ST model provides competitive results comparing to a cascaded model. Our model achieves 0.6 and 0.5 BLEU score improvement in En-Fr and En-Es datasets respectively. The results illustrate that our mutual-learning paradigm provides an effective method for leveraging the additional information available via transcript.

Method	En-Fr	En-Es
Cascaded	34.9	28.0
E2E	32.8	27.2
E2E + MTL	33.5	27.5
E2E + KD	34.5	27.9
E2E + ML	35.5	28.5
E2E + ML*	**36.3**	**28.7**

Table 1: A comparison of ST task evaluation results for different approaches: cascaded model, vanilla end-to-end, end-to-end with multi-task learning, end-to-end with knowledge distillation, and end-to-end with mutual-learning (ML). "*" denotes training with a joint vocabulary.

4.2 Comparison with a Knowledge Distillation Model

Knowledge distillation (KD) is a conceptually similar approach to the proposed framework. KD provides a one way transfer from a trained teacher model to a student model. We provide a focused comparison with this method: we pre-train an MT model using input language and target language sentences, freeze it and use it to guide an ST model by minimising Eq. 13:

$$Loss = \beta * (KL_1 + KL_2) + LC \qquad (13)$$

where KL_1 and KL_2 are described by Eqs. 4 - 5 and LC is the reconstruction loss (Eq. 6). The main difference between KD and our approach is that the MT model is pre-trained, frozen and used in inference mode only to guide the ST model training, which is performed separately form MT model training. From table 1 it can be seen that the proposed mutual-learning approach outperforms one way knowledge distillation strategy (E2E+KD) by 1.0 and 0.6 BLEU score on the En-Fr and En-Es datasets, respectively.

4.3 Comparison with a Multi-Task Learning Model

Multi-Task Learning (MTL) is also a collaborative learning strategy. In contrast to the proposed mutual-learning scenario, in MTL we train all tasks in parallel: ST model and MT model are trained separately with the average of the NLL loss from MT and ST models:

$$Loss = \frac{1}{2} * (LC_{st} + LC_{mt}) \qquad (14)$$

Evaluation of ST task after training using the MTL strategy is shown in Table 1 (E2E + MTL). These results show that our mutual-learning strategy is a more effective way of joint learning: gaining 0.7, 0.3 BLEU score increase over MTL in ST task.

4.4 Joint vocabulary training

Vanilla E2E ST model uses separate vocabularies for source and target languages. We also utilised a jointly-trained byte pair encoding (BPE) to build the vocabulary and achieved a surprising improvement on what was already a state-of-the-art result (see the last row of Table 1).

4.5 Evaluation of the MT task

Method	En-Fr	En-Es
MT	45.1	35.4
MT+MTL	45.6	35.3
MT+ML	**45.8**	**35.7**

Table 2: A comparison of MT task evaluation results for different approaches: independently-trained, multi-task learning and the proposed mutual-learning.

In addition, we evaluate the performance of the MT task and compare our proposed mutual-learning scenario with an independently trained MT model and also the multi-task learning scenario. The architecture of MT model and the hyper-parameters' values used in each training scenario are identical, as described in Sec.3.2.

From the results in Table 2 we can conclude that mutual-learning also improves the MT model's performance. Our system gains 0.7 and 0.3 BLEU score in En-Fr and En-Es datasets, respectively, compared to the independently trained MT system. Our system also exceeds a typical MTL approach by 0.2 and 0.4 BLEU score in the MT task. These results suggest that our mutual-learning leads to a more robust minima than the MTL paradigm.

5 Conclusion

We proposed a mutual-learning paradigm for end-to-end speech translation to effectively transfer knowledge between ST and MT models. Experimental results demonstrate that our proposed approach outperforms knowledge distillation, the typical one-way transfer paradigm, as well as, multi-task learning, a typical dual knowledge transfer paradigm. We also provide a competitive result compared to a cascaded model, which has thus far been outperforming E2E ST models.

References

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 474–479, Valencia, Spain. Association for Computational Linguistics.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, abs/1612.01744.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019b. Adapting transformer to end-to-end spoken language translation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 1133–1137. International Speech Communication Association.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California. Association for Computational Linguistics.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020. On knowledge distillation for direct speech translation. *ArXiv*, abs/2012.04964.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 1128–1132. International Speech Communication Association.

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstra-*

tions(NAACL), pages 82–86, San Diego, California. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Ashutosh Saboo and Timo Baumann. 2019. Integration of dubbing constraints into machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 94–101, Florence, Italy. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Laura Cross Vila, Carlos Escolano, José A. R. Fonollosa, and M. Costa-jussà. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. 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, Suzhou, China. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wuand, and Zhifeng Chen. 2017. Sequence to sequence models can directly transcribe foreign speech. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2625–2629. International Speech Communication Association.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019. Lattice transformer for speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6475–6484, Florence, Italy. Association for Computational Linguistics.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 4320–4328.

Haojie Zhao, Gang Yang, Dong Wang, and Huchuan Lu. 2021. Deep mutual learning for visual object tracking. *Pattern Recognition*, 112:107796.