CKDST: Comprehensively and Effectively Distill Knowledge from Machine Translation to End-to-End Speech Translation

Yikun Lei[1], Zhengshan Xue[1], Haoran Sun[1], Xiaohu Zhao[1], Shaolin Zhu[1]
Xiaodong Lin[3], Deyi Xiong[1,2]*
[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] School of Computer Science and Technology, Kashi University, Kashi, China
[3] Department of Management Science and Information Systems, Rutgers University
{yikunlei,xuezhengshan,hrsun,xhzhao,slzhu,dyxiong}@tju.edu.cn
lin@business.rutgers.edu

Abstract

Distilling knowledge from a high-resource task, e.g., machine translation, is an effective way to alleviate the data scarcity problem of end-to-end speech translation. However, previous works simply use the classical knowledge distillation that does not allow for adequate transfer of knowledge from machine translation. In this paper, we propose a comprehensive knowledge distillation framework for speech translation, **CKDST**, which is capable of comprehensively and effectively distilling knowledge from machine translation to speech translation from two perspectives: cross-modal contrastive representation distillation and simultaneous decoupled knowledge distillation. In the former, we leverage a contrastive learning objective to optimize the mutual information between speech and text representations for representation distillation in the encoder. In the later, we decouple the non-target class knowledge from target class knowledge for logits distillation in the decoder. Experiments on the MuST-C benchmark dataset demonstrate that our CKDST substantially improves the baseline by 1.2 BLEU on average in all translation directions, and outperforms previous state-of-the-art end-to-end and cascaded speech translation models. The source code is available at https://github.com/ethanyklei/CKDST.

1 Introduction

End-to-end (E2E) speech-to-text translation (ST), directly translating speech in one language into text in another, has recently attracted increasing attention (Duong et al., 2016; Zhang et al., 2020; Xu et al., 2021; Ye et al., 2022). Compared with traditional cascaded ST, E2E ST does not require automatic transcription, which endows itself with less error propagation and lower latency.

However, parallel ST data that consist of speech inputs and target translations, are proverbially limited, especially in comparison with automatic speech recognition (ASR) and machine translation (MT) data. In order to mitigate this issue, previous efforts leverage pre-training approaches (Xu et al., 2021; Ao et al., 2022) and multi-task learning (MTL) frameworks (Ye et al., 2021; Tang et al., 2021; Han et al., 2021) to transfer knowledge from ASR and/or MT to ST. Among them, knowledge distillation (KD) (Hinton et al., 2015) has proved to be an effective way to improve ST performance by transferring knowledge from MT to ST (Liu et al., 2019; Xu et al., 2021; Tang et al., 2021).

However, previous KD approaches to ST only explore the classical KD that transfers knowledge from prediction logits, which may not allow for sufficient knowledge distillation. Specifically, in classical KD (Hinton et al., 2015), two types of knowledge are encoded in prediction logits, target class knowledge from target class logits and non-target class knowledge from non-target class logits. Each type of knowledge contributes to the success of classical logits distillation. However, Zhao et al. (2022) have found that the classical KD couples the non-target class knowledge with the target class knowledge. Such entanglement may inhibit the transfer of non-target class knowledge and limit the performance of logits knowledge distillation.

Additionally, due to the modality gap between speech and text, it might be difficult for E2E ST to sufficiently capture and translate semantic information embedded in speech inputs to target translations. Fortunately, however, in MTL-based E2E ST, a speech input is accompanied with its transcription that is used as the input fed into MT. Such speech and transcription pairs allow us to distill knowledge from transcription representations to speech representations so as to reduce the modality gap. However, such knowledge distillation has not yet been explored for end-to-end speech translation.

In order to address these two issues and efficiently distill MT knowledge to ST, we propose

*corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 3123–3137

a **C**omprehensive **K**nowledge **D**istillation framework for **ST** (**CKDST**). Specifically, we propose Cross-modal Contrastive Representation Distillation (CCRD) and Simultaneous Decoupled Knowledge Distillation (SDKD) as two essential approaches for CKDST, to transfering knowledge from text representations and to performing more sufficient logits distillation.

CCRD applies a contrastive training objective to force E2E ST to learn speech representations that are closer to their corresponding textual representations. In doing so, we could increase the mutual information lower bound between speech and text representations (Tian et al., 2019). **SDKD** is proposed for E2E ST to mitigate the issue that the classical KD couples the target class knowledge with non-target class knowledge (Zhao et al., 2022). For more effectively transferring logits knowledge from MT to ST, we decouple these two types of knowledge in prediction logits and extend the decoupled knowledge distillation to the MTL framework where both ST and MT are fine-tuned simultaneously.

In a nutshell, our contributions are three-fold.

- We propose CKDST for end-to-end ST, which can comprehensively and effectively transfer MT knowledge to ST in both the encoder and decoder.
- We introduce CCRD and SDKD in CKDST to increase the mutual information between speech and text representations, and to decouple the non-target class knowledge from the target knowledge for more effective logits distillation, respectively.
- We conduct extensive experiments on the MuST-C benchmark dataset with four language pairs. Experiment results validate the effectiveness of the two approaches and demonstrate that our model outperforms previous best end-to-end and cascaded baselines.

2 Related Work

End-to-End Speech Translation. To alleviate the error propagation in cascaded ST and to ease the deployment, Bérard et al. (2016) and Weiss et al. (2017) propose to use an end-to-end architecture to directly translate speech in one language into text in another, without using the intermediate transcriptions. In recent years, increasing efforts have been done in E2E ST (Di Gangi et al., 2019b; Liu et al., 2019; Wang et al., 2020b; Liu et al., 2020; Xu et al., 2021; Tang et al., 2021; Fang et al., 2022; Tang et al., 2022). Since the parallel speech translation data is notoriously limited, many approaches have been proposed to solve this problem, such as pre-training (Wang et al., 2020b; Xu et al., 2021; Tang et al., 2022), multi-task learning (Le et al., 2020; Zhao et al., 2021; Ye et al., 2022), and data augmentation (Bahar et al., 2019; Lam et al., 2022). Additionally, knowledge distillation from a well trained MT model to a ST model has proved effective in improving ST performance. Liu et al. (2019) leverage knowledge distillation to allow the E2E ST model to learn the same prediction distribution as the MT model. The MT model is frozen while the ST model is being trained. SATE (Xu et al., 2021) uses both pre-trained ASR model and MT model as teacher models to perform knowledge distillation. Each pre-trained model serves a different module of the ST model, and they are also frozen during training. Tang et al. (2021) propose the online-KD that simultaneously update the ST module and the MT module in a multi-task learning framework. However, these efforts only distill the knowledge from prediction logits via classical KD, and the knowledge from encoder representations is ignored. In our work, we comprehensively and efficiently distill knowledge from both encoder representations and prediction logits of MT to ST.

Knowledge Distillation. The concept of knowledge distillation has been firstly proposed by Hinton et al. (2015). KD defines a learning framework where a stronger teacher network is employed to guide the training of a student network for many tasks (Kim and Rush, 2016; Li et al., 2017; Tan et al., 2019). The subsequent works can be roughly divided into two groups, distillation from prediction logits (Furlanello et al., 2018; Cho and Hariharan, 2019; Yang et al., 2019; Mirzadeh et al., 2020) and intermediate representations (Yim et al., 2017; Huang and Wang, 2017; Heo et al., 2019; Park et al., 2019). Romero et al. (2014) explore intermediate representations for KD by using regressions to guide the feature activations of the student network. Tian et al. (2019) apply a contrastive objective to maximize the mutual information lower bound between teacher representations and student representations. In contrast, DKD (Zhao et al., 2022) decouples and amplifies student-friendly knowledge from prediction distribution to perform more

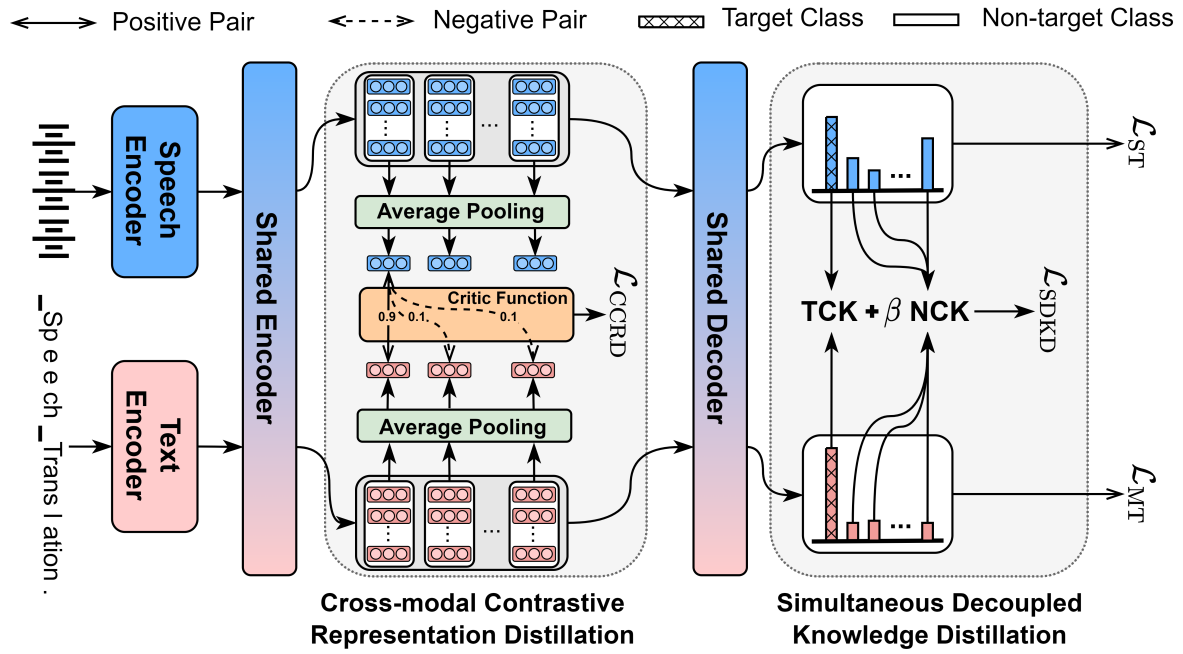


Figure 1: The diagram of CKDST. The blue part is related to the ST task while the light pink part is related to the MT task. "Average Pooling" is applied on the time dimension to obtain the sentence-level representations.

effective distillation. Our approaches are partially motivated by these previous efforts but are significantly different from them in two aspects. First, the knowledge gap in representations is due to different modalities rather than the model capability (teacher vs. student), so the inconsistency of modalities is also an important challenge we have to deal with in our distillation. Second, we don't freeze the teacher during distillation, we argue that this may allow the teacher model to adapt to provide student-friendly knowledge.

3 CKDST

In this section, we first introduce the model architecture of CKDST and then elaborate the two knowledge distillation approaches in CKDST.

3.1 Model Architecture

CKDST adopts the encoder-decoder ST framework, as shown in Figure 1. It consists of four main components: speech encoder, text encoder, shared encoder and shared decoder, facilitating the joint training of the ST and MT task.

Speech Encoder is composed of non-finetuned wav2vec 2.0 (Baevski et al., 2020) followed by two layers of 1-D CNNs. It takes speech wavforms as input to obtain low-level speech representations.

Text Encoder is the normal word embedding layer, which is the same as the word embedding layer for text translation. It takes text as input for the MT task.

Shared Encoder / Decoder adopt the standard Transformer (Vaswani et al., 2017) as their backbone network. The shared encoder takes outputs from both speech and text encoder as inputs to further extract semantic information. The shared decoder generates target translations for ST and MT. And, with shared parameters, the shared encoder and decoder are expected to learn the shared knowledge between ST and MT.

A training sample for E2E ST is a *(speech,transcript,translation)* triplet (s, t, y). We use speech-translation pairs (s, y) as training data for ST, and transcript-translation pairs (t, y) as training data for MT. The cross-entropy loss is adopted for both ST and MT:

$$\mathcal{L}_{\text{ST}} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, s)$$
$$\mathcal{L}_{\text{MT}} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, t) \quad (1)$$

3.2 Cross-modal Constrative Representation Distillation

Speech inputs are usually noisier than their textual counterpart transcripts (Tang et al., 2021), which makes the extraction of semantic information from

speech difficult. Thus, we want to transfer semantic knowledge across modality, from text representations to speech representations. For this, we propose cross-modal contrastive representation distillation which employs a contrastive training objective to maximize the mutual information between text and speech representations. Due to the length difference between speech and text, we use sentence-level representations for distillation. We apply average pooling on the output of the shared encoder in the time dimension to obtain sentence-level representations of speech and text.

Concretely, let $\boldsymbol{T}$ and $\boldsymbol{S}$ denote the sentence-level source text representation (teacher representation) in MT and speech representation (student representation) in ST, respectively. Tian et al. (2019) show that a lower bound of the mutual information (MI) between $\boldsymbol{T}$ and $\boldsymbol{S}$ exists when we have 1 positive pair (i.e., speech-transcript pair) and N negative pairs (i.e., pairs of a speech with the rest of the transcripts in the same mini batch). The lower bound is estimated as follows:

$$\begin{aligned}\mathrm{MI}(\boldsymbol{T},\boldsymbol{S}) \geq \log(N) + \mathbb{E}_{q(\boldsymbol{T},\boldsymbol{S})}[\log h(\boldsymbol{T},\boldsymbol{S})] \\ + N\mathbb{E}_{q(\boldsymbol{T})q(\boldsymbol{S})}[\log\left(1-h(\boldsymbol{T},\boldsymbol{S})\right)]\end{aligned} \tag{2}$$

where $q(\boldsymbol{T},\boldsymbol{S})$ (indicating positive pairs) and $q(\boldsymbol{T})q(\boldsymbol{S})$ (indicating negative pairs) are the joint distribution and the product of marginal distributions of $\boldsymbol{T}$ and $\boldsymbol{S}$, respectively. $h(\cdot)$ is a critic function that estimates the probability that the input pair $(\boldsymbol{T},\boldsymbol{S})$ is drawn from $q(\boldsymbol{T},\boldsymbol{S})$.

We optimize our model to maximize the expectation terms in Eq. (2). In doing so, we force our model to learn a representation for $\boldsymbol{S}$, which is semantically close to that of $\boldsymbol{T}$, so as to optimize the mutual information between $\boldsymbol{S}$ and $\boldsymbol{T}$. This can be considered as a procedure that distills knowledge from the representation of $\boldsymbol{T}$ to that of $\boldsymbol{S}$.

Given a sentence-level text-speech representation pair $(\boldsymbol{T},\boldsymbol{S})$, the critic function $h(\cdot)$ calculates a score indicating the possiblity that $(\boldsymbol{T},\boldsymbol{S})$ is a positive pair (drawn from $q(\boldsymbol{T},\boldsymbol{S})$) or negative pair (drawn from $q(\boldsymbol{T})q(\boldsymbol{S})$). We define the critic function $h(\boldsymbol{T},\boldsymbol{S}) \rightarrow [0,1]$ as follows:

$$h(\boldsymbol{T},\boldsymbol{S}) = \frac{e^{cos(\boldsymbol{T},\boldsymbol{S})/\tau}}{1+e^{cos(\boldsymbol{T},\boldsymbol{S})/\tau}} \tag{3}$$

where $cos(\cdot)$ is the cosine similarity and τ is a temperature hyper-parameter. With the critic probability $h(\boldsymbol{T},\boldsymbol{S})$, we can calculate the loss of CCRD:

$$\begin{aligned}\mathcal{L}_{\mathrm{CCRD}} = -\left(\mathbb{E}_{q(\boldsymbol{T},\boldsymbol{S})}[\log h(\boldsymbol{T},\boldsymbol{S})]\right. \\ \left. + N\mathbb{E}_{q(\boldsymbol{T})q(\boldsymbol{S})}[\log\left(1-h(\boldsymbol{T},\boldsymbol{S})\right)]\right)\end{aligned} \tag{4}$$

Different from Ye et al. (2022) who use contrastive learning to bridge the modality gap between speech and text, we use NCE loss (Gutmann and Hyvärinen, 2010) as the contrastive objective, and aim to maximize the lower bound of mutual information between speech and text representations.

3.3 Simultaneous Decoupled Knowledge Distillation

Previous efforts (Liu et al., 2019; Xu et al., 2021; Tang et al., 2021) calculate the KL-Divergence on prediction logits between ST and MT for logits distillation. However, this classical KD loss (Hinton et al., 2015) couples target class knowledge and non-target class knowledge by the confidence of the teacher model on the target class, and suppress the non-target class knowledge transfer which limits the effectiveness of logits distillation (Zhao et al., 2022). Therefore, we propose simultaneous decoupled knowledge distillation which decouples the non-target class knowledge from target class knowledge to allow more sufficient knowledge distillation than the classical KD.

Let p_i^T and p_i^S be the probabilities of MT and ST for the i-th subword in the vocabulary V, respectively. The classical KD loss can be formulated as:

$$\mathcal{L}_{\mathrm{KD}} = \sum_{i=1}^{|V|} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right) \tag{5}$$

We use p_t to denote the probability of the target subword. Correspondingly, the sum of the probabilities of the remaining non-target subwords is $p_{\setminus t} = (1-p_t)$. Meanwhile, let $\hat{p}_i$ be the probability of modeling on non-target subwords (i.e., without considering the target class), which can be calculated as:

$$\hat{p}_i = \frac{\exp(z_i)}{\sum_{j=1,j\neq t}^{|V|}\exp(z_j)}(i \neq t) \tag{6}$$

where z is the logit. Since $\hat{p}_i$ is independent of the target subword probability p_t, we assume that it represents the non-target class knowledge in prediction logits. Now, according to the above definitions,

we can reformulate Eq. (5) as:

$$\mathcal{L}_{KD} = \underbrace{p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + p_{\backslash t}^T \log\left(\frac{p_{\backslash t}^T}{p_{\backslash t}^S}\right)}_{\textbf{TCK}} + (1-p_t^T) \underbrace{\sum_{i=1, i\neq t}^{|V|} \hat{p}_i^T \log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right)}_{\textbf{NCK}} \quad (7)$$

The details of the reformulation can be found in Appendix A. Obviously, the non-target class knowledge (NCK) couples with the target class knowledge (TCK) with a coupling weight $(1-p_t^T)$. Thus larger prediction scores p_t^T of the teacher model would lead to smaller coupling weights of NCK, which significantly suppresses the transfer of NCK. However, such suppression is not desirable since the more confident the teacher is, the more reliable and valuable knowledge it can provide, and since the contributions of NCK and TCK are from different aspects that should be considered separately (Zhao et al., 2022). Therefore, we replace $(1-p_t^T)$ with a hyper-parameter β to decouple the TCK and NCK to control the importance of the two types of knowledge, separately, and the training objective of SDKD is calculated as follows[1]:

$$\mathcal{L}_{\text{SDKD}} = \textbf{TCK} + \beta\textbf{NCK} \quad (8)$$

3.4 Training and Inference

We train our model in a pretraining-then-finetuning manner. We first pre-train the text encoder, shared encoder and decoder with MT data. Then during the fine-tuning phase, we jointly train ST, MT, CCRD and SDKD with ST data. The overall training objective is the combination of the four task losses:

$$\mathcal{L} = \mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{MT}} + \mathcal{L}_{\text{CCRD}} + \mathcal{L}_{\text{SDKD}} \quad (9)$$

Note that we do not freeze MT parameters (i.e., we still enable the gradient propagation of MT) when distilling knowledge from representations and prediction logits. This is because we find that contiguously training MT parameters benefits ST performance in our experiments (see Appendix B).

During inference, we remove the text encoder and use the remaining modules of CKDST for speech translation.

[1]We regard this as decoupling because the knowledge transfer of NCK is not controlled (i.e., weighted) by the probability of the target class any more. Instead, it is controlled by β.

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

Table 1: Statistics of the used datasets.

4 Experiments

We compared with state-of-the-art E2E/cascaded ST models to examine the effectiveness of the proposed CKDST.

4.1 Datasets

ST Dataset We conducted experiments on the MuST-C[2] (Di Gangi et al., 2019a) benchmark dataset in four translation directions: English-German (En-De), English-Spanish (En-Es), English-French (En-Fr) and English-Russian (En-Ru). Each direction have around 400 hours speech. *dev* was used to develop and analyze our approaches, *tst-common* was used for testing.

External MT Data We followed previous works (Tang et al., 2021; Ye et al., 2022) to use WMT[3] datasets of different years as external MT data: WMT 2016 for English-German and English-Russian, WMT 2014 for English-French and WMT 2013 for English-Spanish.

The statistics of MuST-C and WMT datasets are shown in Table 1.

4.2 Settings

Pre-processing We used 16-bit 16 kHz mono-channel audio wave as speech input. And we removed utterances of which the duration is longer than 30s. For text inputs, we extracted 10K unigram subwords with a shared source and target vocabulary via SentencePiece[4] (Kudo and Richardson, 2018).

Model Configuration We used the base version of Wav2vec 2.0[5] in the speech encoder, which is pretrained on audio data from LibriSpeech (Panayotov et al., 2015) without finetuning. Two layers of CNNs were stacked over Wav2vec 2.0 , where the kernel size was set to 5, stride size to 2 and hidden

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

[3]https://statmt.org/

[4]https://github.com/google/sentencepiece

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

Models	External Data			MuST-C				
	Speech	ASR	MT	En-De	En-Es	En-Fr	En-Ru	AVG
w/o external MT data								
Fairseq ST (Wang et al., 2020a)	-	-	-	22.7	27.2	32.9	15.3	24.5
Espnet ST (Inaguma et al., 2020)	-	-	-	22.8	27.4	33.3	15.6	24.8
W-Transf (Ye et al., 2021)	✓	-	-	23.6	28.4	34.6	14.4	25.3
XSTNet (Ye et al., 2021)	✓	-	-	25.5	29.6	36.0	16.9	27.0
STEMM (Fang et al., 2022)	✓	-	-	25.6	30.3	36.1	17.1	27.3
ConST (Ye et al., 2022)	✓	-	-	25.7	30.4	36.8	17.3	27.6
MTL baseline	✓	-	-	25.4	29.6	35.9	16.8	26.9
Ours	✓	-	-	**26.4**	**30.9**	**37.3**	**17.7**	**28.1**
w/ external MT data								
JT-S-MT (Tang et al., 2021)	-	-	✓	26.8	31.0	37.4	-	-
SATE (Xu et al., 2021)	-	✓	✓	28.1†	-	-	-	-
Chimera (Han et al., 2021)	✓	-	✓	27.1†	30.6	35.6	17.4	27.7
XSTNet (Ye et al., 2021)	✓	-	✓	27.1	30.8	38.0	18.5	28.6
STEMM (Fang et al., 2022)	✓	-	✓	**28.7**	31.0	37.4	17.8	28.7
ConST (Ye et al., 2022)	✓	-	✓	28.3	32.0	38.3	18.9	29.4
MTL baseline	✓	-	✓	27.1	31.2	37.3	18.2	28.5
Ours	✓	-	✓	28.5	**32.5**	**38.5**	**19.1**	**29.7**

Table 2: BLEU scores of different models on the MuST-C *tst-common* set. "Speech" indicates unlabelled speech data. "MTL baseline" is the implemented strong baseline using the same architecture as our model, excluding CCRD and SDKD. †denotes that large-scale Opensubtitles (Lison and Tiedemann, 2016) data are used as the external MT data.

size to 512. The shared encoder and decoder were configured with the base Transformer setting: 6 layers, 512 as hidden size, 8 attention heads, and 2048 as FFN hidden size.

Implementation Details We implemented our model based on fairseq toolkit.[6] For experiments with external data, we used external MT data for pre-training. For those without any external data, only the MT data from the ST triplet data were considered. For fine-tuning, we used the same hyper-parameters for experiments with/without external MT data. Particularly, we used the Adam optimizer wiht 25K warm-up updates. The learning rate was 1e-4. The maximal number of tokens was 0.8M per batch. Both the dropout and the value of label smoothing were set to 0.1. We set the update frequency to 2. The temperature τ was 0.1 and the non-target class knowledge weight β was 4.0. We set the maximal number of updates to 200000, and used the early-stop training strategy if the performance did not improve for 10 consecutive validation runs. We trained all the models on 4 Nvidia TeslaV100 GPUs.

During inference, we averaged the checkpoints of the last 10 epochs for evaluation. We used beam search with a beam size of 10 and length penalty was 1.0. We evaluated case-sensitive detokenized BLEU and ChrF++ by sacreBLEU[7] (Post, 2018). Additionally, we also evaluated translation quality with COMET (Rei et al., 2020), which leverages pretrained language models to achieve high correlations with human quality judgments. Specifically, we used COMET-22 (wmt22-COMET-da)[8].

Baselines We compared our CKDST with multiple strong E2E ST baselines including: (1) Fairseq ST (Wang et al., 2020a) and (2) Espnet ST (Inaguma et al., 2020) trained only with the ST task data, (3) W-Transf (Ye et al., 2021) that uses a pre-trained speech model to extract speech features, (4) XSTNet (Ye et al., 2021) that trains the ST model based on W-Transf in a multitask learning framework, (5) Chimera (Han et al., 2021) that learns a shared memory space to align speech and text, (6) STEMM (Fang et al., 2022) that mixes speech and text representations and (7) ConST (Ye et al., 2022) that applies contrastive learning to bridge the modality gap between speech and text, (8) JT-S-MT (Tang et al., 2021) that employs an online-KD method to transfer knowledge from MT

[6] https://github.com/facebookresearch/fairseq

[7] sacreBLEU signature: BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+tok.13a+version.1.5.1

[8] https://huggingface.co/Unbabel/wmt22-comet-da

	ChrF++					COMET				
	En-De	**En-Es**	**En-Fr**	**En-Ru**	**AVG**	**En-De**	**En-Es**	**En-Fr**	**En-Ru**	**AVG**
MTL baseline	55.02	57.92	61.71	43.39	54.51	82.15	82.10	81.00	80.34	81.29
Ours	**55.99**	**58.77**	**62.54**	**45.04**	**55.59**	**82.67**	**82.69**	**81.55**	**82.38**	**82.32**

Table 3: Results of ChrF++ and COMET for the four language pairs on the MuST-C benchmark dataset.

Models	**En-De**	**En-Fr**
Cascaded		
Espnet (Inaguma et al., 2020)	23.6	33.8
(Xu et al., 2021)	28.1	-
Cascaded baseline	27.2	36.6
End-to-end		
Ours	**28.5**	**38.5**

Table 4: Comparison to cascaded baselines on the MuST-C En-De and En-Fr *tst-common* set. "Cascaded baseline" is the implemented strong cascaded baseline which uses the speech encoder of our model as its ASR module and the MT related part of our model as its MT module.

to ST and (9) SATE (Xu et al., 2021) that leverages an adapter to incorporate pre-trained ASR and MT models into E2E ST, and uses classical KD for knowledge transfer. In addition to these baselines, we implemented a strong baseline "MTL baseline" that uses the same neural architecture (excluding the proposed CCRD and SDKD) as our model to jointly train ST and MT.

4.3 Main Results

Comparison to End-to-End Baselines. We compared our model with several strong baselines for four language pairs on the MuST-C benchmark dataset. Results are shown in Table 2. Without the external MT data, our model achieves a substantial improvement of 1.2 BLEU over the MTL baseline on average and outperforms the strongest baseline, ConST, in all translation directions. When we use the external MT data, we achieve new state-of-the-art results in terms of the average BLEU score over the four translation directions and gain a 1.2 BLEU improvement over the MTL baseline. These results demonstrate that our approaches are able to effectively improve ST with knowledge distillation. Compared to previous works that explore knowledge distillation for E2E ST, we outperform JT-S-MT (Tang et al., 2021) and SATE (Xu et al., 2021) by 1.4 BLEU and 0.4 BLEU on average, respectively. This suggests that our proposed knowledge distillation approaches are more effective than previous KD methods used in E2E ST. To better evaluate our approach, we used ChrF++ and COMET, which are more relevant to human evaluation, to assess our model. As shown in Table 3, our model achieves an average improvement of 1.08 ChrF++ and 1.03 COMET compared to the MLT baseline model.

Comparison to Cascaded Baselines. We also compared our end-to-end model with cascade baselines. Espnet (Inaguma et al., 2020) and the cascaded ST system presented by Xu et al. (2021) are two strong cascaded systems trained with MuST-C and external ASR and MT data (LibriSpeech, WMT, and Opensubtitles). We implemented a strong "Cascaded baseline" using the ASR data from the ST data and the same external MT data as ours. Its ASR module is the same as our speech encoder and was trained with the CTC loss. The MT module is a standard Transformer, trained with the traditional MT loss. As shown in Table 4, our implemented Cascaded baseline is competitive to the other two cascaded baselines. Impressively, our end-to-end model outperforms all cascaded baselines in all translation directions.

4.4 Ablation Study

To better evaluate the contribution of our proposed knowledge distillation approaches, we progressively removed the CCRD module and the SDKD module to conduct ablation study on the MUST-C benchmark. As shown in Table 5, without CCRD, we get an average drop of 0.5 BLEU on all four translation directions. And, SDKD also contributes 0.6 BLEU on average on all translation directions. These demonstrate the effectiveness of both approaches in enhancing ST.

5 Analysis

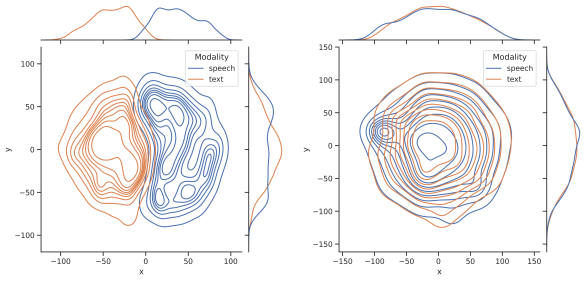
Additionally, we conducted a series of in-depth analyses to further investigate how the proposed methods improve E2E ST.

5.1 Does CCRD Increase the Mutual Information?

The proposed CCRD distills knowledge from MT to ST by optimizing the mutual information be-

Ablation	MuST-C			
	En-De	**En-Es**	**En-Fr**	**En-Ru**
Ours	28.5	32.5	38.5	19.1
- $\mathcal{L}_{\text{CCRD}}$	28.2	32.1	38.2	18.7
- $\mathcal{L}_{\text{SDKD}}$	28.0	31.9	37.9	18.5

Table 5: BLEU scores on MuST-C benchmark *tst-common* set by removing individual losses.



(a) w/o CCRD (b) w/ CCRD

Figure 2: Bivariate KDE contour of the speech and text representations. t-SNE is used to reduce the dimension into 2D. Blue curves are speech representations while orange curves stand for text representations. Samples are drawn from the MuST-C En-De *dev* set.

tween text and speech representations. Mutual information (MI) can be represented by the degree of overlap between two distributions. Thus, we plot the bivariate kernel density estimation (Parzen, 1962) (KDE) contour of speech and text dim-reduced representations to visualize their distributions as shown in Figure 2, where t-SNE (Van der Maaten and Hinton, 2008) is used to reduce the dimension of representations into 2D. As shown in Figure 2(a), without CCRD, the overlap of the speech representation distribution and the text representation distribution is small. This shows that even with the shared encoder, the distributions of representations from the two modalities have very low MI. In contrast, when we apply CCRD, the distribution of speech representations and the distribution of text representations almost overlap. This indicates our proposed CCRD can significantly improve the MI between the two representation distributions.

5.2 Is SDKD Better than Classical KD?

As discussed in Section 3.3, the classical KD suppresses the knowledge of non-target classes, which limits its performance. To verify this, we conducted experiments on the MuST-C benchmark to compare the effects of SDKD and classical KD. In order to

Loss	MuST-C			
	En-De	**En-Es**	**En-Fr**	**En-Ru**
$\mathcal{L}_{\text{SDKD}}$	28.2	32.1	38.2	18.7
$\mathcal{L}_{\text{KD}}$	27.9	31.6	38.1	18.3

Table 6: SDKD vs. the classical KD on the MuST-C benchmark. $\mathcal{L}_{KD}$ is calculated according to Eq. (5).

eliminate the interference of other factors, we did not apply CCRD during training. The loss function $\mathcal{L}_{\text{KD}}$ of the classical KD is estimated according to Eq. (5). During training, $\mathcal{L}_{KD}$ is interpolated with the primary loss (i.e., ST loss) with weight α (Hinton et al., 2015). Therefore, the training objective for E2E ST with the classical KD is:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{ST}} + \alpha\mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{MT}} \quad (10)$$

We followed Tang et al. (2021) to set α to 0.8. As shown in Table 6, SDKD outperforms the classical KD on all translation directions and achieves an average improvement of 0.4 BLEU. This indicates that separately exploring the target and non-target class knowledge is better than the coupled form.

5.3 Impact of the Non-target Class Knowledge Weight

For SDKD, it is important to choose an appropriate non-target class knowledge weight β. To understand the impact of β, we employed a grid search from $[0, 8]$ to search desirable β with a stride of 2 on the MuST-C En-De *dev* set. Results are shown in Figure 3. The orange dashed line indicates the baseline model which uses the classical KD during training. If $\beta = 0$, it indicates that the non-target class knowledge is ignored when distilling knowledge from prediction logits. Compared with the classical KD baseline, the model performance drops significantly if we ignore the non-target class knowledge. This suggests that the non-target class knowledge is important and useful. The curve with varying β clearly shows that the performance of the model first increases and then drops as β increases. We achieve the best BLEU score when $\beta = 4$. This indicates that appropriately increasing the importance of the non-target class knowledge is beneficial for knowledge distillation, but too large weights would undermine the performance of the model.

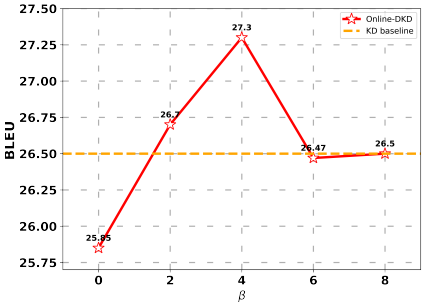


Figure 3: BLEU score curve on the MuST-C En-De *dev* set with the non-target class knowledge weight β varying from 0 to 8. Orange dashed curve indicates the baseline model which uses the classical KD during training. $\beta = 0$ indicates that the non-target class knowledge is ignored when distilling knowledge from prediction logits.

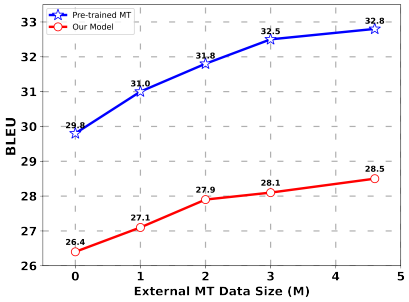


Figure 4: BLEU score curve on the MuST-C En-De *tst-common* against the size of external MT data used during pre-training. When the size of external MT data is 0, we use the MT data from the ST training data for pre-training.

5.4 Impact of the Performance of the Pre-trained MT Model

Our proposed approaches aim to effectively distill knowledge from MT to ST, thus the pre-trained MT performance is of importance to our model. In order to study the impact of MT performance on our model, we randomly sample 1M, 2M and 3M MT data from the external MT data to pre-train the MT model so as to have different MT models with varying performance. When the size of external MT data is 0, we use the MT data from the ST triplet to pre-train the MT model. Results are shown in Figure 4. We observe that as the performance of the pre-trained MT model improves, the BLEU score of our model also keeps improving. This demonstrates that our approaches benefit from strong pre-trained MT models.

6 Conclusion

In this paper, we have presented CKDST, which comprehensively and effectively distills the knowledge of MT to boost the performance of E2E ST through two key approaches: CCRD and SDKD. The former leverages a contrastive objective to maximize the mutual information lower bound between speech and text representations for representation knowledge distillation. The later reformulates the classical KD loss to decouple the target class knowledge and the non-target class knowledge for more effective logits knowledge distillation. Our experiments strongly demonstrate that our approaches are able to significantly improve E2E ST and achieve new state-of-the-art results on the MUST-C benchmark dataset.

Limitations

Although the proposed CKDST distills the knowledge of MT more comprehensively and efficiently from encoder representations and prediction logits, and obtains significant improvements over previous methods, it still has limitations: (1) The batch size is not very large, limited by the memory capacity of the used hardware and the extremely long sequence length of speech inputs, which leads to a small number of negative samples used in CCRD and does not fully exploit the ability of contrastive learning. In future work, we attempt to expand the negative sample size using a mechanism like memory bank (He et al., 2020). (2) As we distill knowledge from MT to ST, the performance of the pretrained MT model has an impact on our framework.

Ethics Statement

This work presents CKDST, a knowledge distillation framework for ST to more comprehensively and effectively distill knowledge from MT to improve the performance of E2E ST. The datasets used in this study include both MuST-C and WMT. They are all public datasets and are widely used in the MT community.

Acknowledgments

The present research was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01D43). We would like to thank the anonymous reviewers for their insightful comments.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

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.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On using specaugment for end-to-end speech translation. *arXiv preprint arXiv:1911.08876*.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019b. One-to-many multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592. IEEE.

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.

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, Dublin, Ireland. Association for Computational Linguistics.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

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, Online. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Zehao Huang and Naiyan Wang. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. 2020. Espnet-st: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

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: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 245–254, Dublin, Ireland. Association for Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition

and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Quanquan Li, Shengying Jin, and Junjie Yan. 2017. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*.

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

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

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)*, pages 5206–5210. IEEE.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

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, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9161–9168.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

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, Online. Association for Computational Linguistics.

Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. 2019. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*.

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, Seattle, United States. Association for Computational Linguistics.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2533–2544, Online. Association for Computational Linguistics.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962.

Jiawei Zhao, Wei Luo, Boxing Chen, and Andrew Gilman. 2021. Mutual-learning improves end-to-end speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3989–3994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Reformulation Details

In Sec 3.3, we define the classical KD loss as follows:

$$\begin{aligned}\mathcal{L}_{KD} &= \sum_{i=1}^{|V|} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right) \\ &= p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + \sum_{i=1,i\neq t}^{|V|} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right)\end{aligned} \tag{11}$$

According to the definition of $p_{\setminus t}$ and $\hat{p}_i$ in Sec 3.3, we can reformulate Eq. (11) to:

$$\begin{aligned}\mathcal{L}_{KD} =& p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + \sum_{i=1,i\neq t}^{|V|} p_{\setminus t}^T \hat{p}_i^T \log\left(\frac{p_{\setminus t}^T \hat{p}_i^T}{p_{\setminus t}^S \hat{p}_i^S}\right) \\ =& p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) \\ &+ \sum_{i=1,i\neq t}^{|V|} p_{\setminus t}^T \hat{p}_i^T \left(\log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right) + \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right)\right) \\ =& p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + \sum_{i=1,i\neq t}^{|V|} p_{\setminus t}^T \hat{p}_i^T \log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right) \\ &+ \sum_{i=1,i\neq t}^{|V|} p_{\setminus t}^T \hat{p}_i^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right)\end{aligned} \tag{12}$$

Since $p_{\setminus t}^T$ and $p_{\setminus t}^S$ are irrelevant to the class index i, we have:

$$\sum_{i=1,i\neq t}^{|V|} p_{\setminus t}^T \hat{p}_i^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right) = p_{\setminus t}^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right) \sum_{i=1,i\neq t}^{|V|} \hat{p}_i^T \tag{13}$$

Moreover, $\sum_{i=1,i\neq t}^{|V|} \hat{p}_i^T = 1$, so:

$$\sum_{i=1,i\neq t}^{|V|} p_{\setminus t}^T \hat{p}_i^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right) = p_{\setminus t}^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right) \tag{14}$$

Bringing Eq. (14) back to Eq. (12), we have:

$$\begin{aligned}\mathcal{L}_{KD} =& p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + p_{\setminus t}^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right) \\ &+ p_{\setminus t}^T \sum_{i=1,i\neq t}^{|V|} \hat{p}_i^T \log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right) \\ =& p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + p_{\setminus t}^T \log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right) \\ &+ (1 - p_t^T) \sum_{i=1,i\neq t}^{|V|} \hat{p}_i^T \log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right)\end{aligned} \tag{15}$$

B Whether to freeze MT Parameters

Knowledge distillation usually freezes the teacher model (i.e., the gradient propagation of the teacher model is disable). We assume that this is because the teacher model is not supervised by the primary loss and freezing the teacher model prevents it from being degraded by the student model. However, our model is trained on ST and MT, simultaneously. The teacher knowledge can be preserved by the auxiliary MT task. Moreover, we assume that not freezing teacher knowledge during knowledge distillation can make it more student-friendly. To investigate this, we conducted experiments on Must-C En-De. Results are shown in Table 7. When we freeze MT in CCRD, we find the performance drops 0.3 BLEU. In SDKD, there is no difference in the performance of freezing MT or not. In general, not freezing MT when performing knowledge distillation is more suitable for our model.

Methods	MT	MuST-C En-De
CCRD	✓	27.7
	✗	28.0
SDKD	✓	28.2
	✗	28.2

Table 7: Freezing MT vs. not freezeing MT on the MuST-C En-De *tst-common* set. ✓indicates freezing MT while ✗ indicates not freezing MT.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
limitations

☑ A2. Did you discuss any potential risks of your work?
limitations

☑ 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?

2

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix D

☑ 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)?
3.2

☑ 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?
3.2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix C

☑ 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.
Appendix C

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?
No response.

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?
No response.

☐ 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?
No response.

☐ 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.)?
No response.

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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.