An Empirical Study of Consistency Regularization for End-to-End Speech-to-Text Translation

Anonymous ACL submission

Abstract

Consistency regularization methods, such as R-Drop (Liang et al., 2021) and CrossConST (Gao et al., 2023), have achieved impressive supervised and zero-shot performance in the neural machine translation (NMT) field. Can we also boost end-to-end (E2E) speech-to-text translation (ST) by leveraging consistency regularization? In this paper, we conduct empirical studies on intra-modal and cross-modal consistency and propose two training strategies, SimRegCR and SimZeroCR, for E2E ST in regular and zero-shot scenarios. Experiments on the MuST-C benchmark show that our approaches achieve state-of-the-art (SOTA) performance in most translation directions. The analyses prove that regularization brought by the intra-modal consistency, instead of modality gap, is crucial for the regular E2E ST, and the cross-modal consistency could close the modality gap and boost the zero-shot E2E ST performance.

1 Introduction

Speech-to-text translation takes acoustic speech signals as input and outputs text translations in the target language. The conventional cascaded ST system consists of an automatic speech recognition (ASR) system and a machine translation (MT) module in a pipeline manner (Sperber et al., 2017, 2019; Zhang et al., 2019). Recent works on ST have focused on the end-to-end system, which learns a unified model that directly generates text translations from speech without any intermediate outputs (Duong et al., 2016; Berard et al., 2016). E2E ST is a cross-modal task, where the major challenges include parallel ST data scarcity and representation discrepancy between speech and text modalities. In order to boost E2E ST training, the techniques utilized by existing approaches include pretraining (Wang et al., 2020b; Xu et al., 2021), multi-task learning (Ye et al., 2021; Tang et al., 2021a), knowledge distillation (Liu et al., 2019; Inaguma et al., 2021), and cross-modal representation learning (Ye et al., 2022; Wang et al., 2022; Fang and Feng, 2023b). However, most methods are far from being widely used due to the sophisticated model architecture, complicated algorithm implementation, and tedious hyperparameter search.

Consistency regularization has been widely adopted and shown great promise to improve NMT performance (Sato et al., 2019; Chen et al., 2021; Liang et al., 2021; Gao et al., 2022, 2023). Specifically, Liang et al. (2021) introduce an intra-lingual consistency, R-Drop, to regularize dropout and improve the supervised NMT performance, and Gao et al. (2023) propose a cross-lingual consistency, CrossConST, to learn universal representations and boost the zero-shot NMT performance. Given the similar problem formulations between NMT and E2E ST, a natural question arises: *Can we significantly improve E2E ST performance by leveraging simple consistency regularization?*

In this paper, our primary goal is to provide a simple, easy-to-reproduce, but tough-to-beat strategy for learning E2E ST models. Inspired by Liang et al. (2021) and Gao et al. (2023), we propose two strategies, SimRegCR and SimZeroCR, for training E2E ST models in regular and zero-shot scenarios. We show that intra-modal consistency is crucial for the regular setting, and cross-modal consistency is the key for closing the modality gap and boosting the zero-shot performance. The contributions of this paper can be summarized as follows:

- We conduct empirical studies on consistency regularization and propose two simple but effective strategies for learning E2E ST models in regular and zero-shot scenarios.

- Experimental results show that our approaches achieve significant improvements on the MuST-C benchmark and outperform the current SOTA methods CRESS (Fang and Feng, 2023b) and DCMA (Wang et al., 2022).

2 Background

2.1 End-to-End Speech-to-Text Translation

Speech translation corpora usually contain speech-transcription-translation triples, which can be denoted as $\mathcal{S} = \{\mathbf{s}^i, \mathbf{x}^i, \mathbf{y}^i\}_{i=1}^{|\mathcal{S}|}$. $\mathbf{s}$ denotes the sequence of the audio wave, $\mathbf{x}$ is the transcription in the source language, and $\mathbf{y}$ represents the translation in the target language. $\mathcal{S}$ could be pairwise combined into three parallel corpora, $\mathcal{S}_{asr} = \{\mathbf{s}^i, \mathbf{x}^i\}_{i=1}^{|\mathcal{S}|}$, $\mathcal{S}_{mt} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^{|\mathcal{S}|}$, and $\mathcal{S}_{st} = \{\mathbf{s}^i, \mathbf{y}^i\}_{i=1}^{|\mathcal{S}|}$, for ASR, MT, and ST tasks respectively. The goal of E2E ST is to generate translation $\mathbf{y}$ directly from the speech $\mathbf{s}$ without generating transcription $\mathbf{x}$, and the standard training objective is to minimize the empirical risk:

$$\mathcal{L}_{ce}^{st}(\theta) = \ell(f(\mathbf{s}, \mathbf{y}; \theta), \ddot{\mathbf{y}}), \tag{1}$$

where ℓ denotes the cross-entropy loss, θ is a set of model parameters, $f(\mathbf{s}, \mathbf{y}; \theta)$ is a sequence of probability predictions, and $\ddot{\mathbf{y}}$ is a sequence of one-hot label vectors for $\mathbf{y}$. Directly modeling the speech-to-text mapping is nontrivial due to the representation discrepancy between speech and text modalities. To alleviate ST data sparsity, people usually include ASR and MT supervisions from $\mathcal{S}_{asr}$ and $\mathcal{S}_{mt}$ as well as external corpora for E2E ST task.

2.2 Consistency Regularization for Neural Machine Translation

Liang et al. (2021) propose an intra-lingual consistency regularization, R-Drop, for boosting NMT performance by forcing the output distributions of different sub-models generated by dropout to be consistent with each other. For each sentence pair $(\mathbf{x}, \mathbf{y})$, the training objective is defined as:

$$\mathcal{L}_{R-Drop}(\theta) = \mathcal{L}_{ce}^{mt}(\theta) + \alpha \mathcal{L}_{intra}^{mt}(\theta), \tag{2}$$

where

$$\mathcal{L}_{ce}^{mt}(\theta) = \ell(f(\mathbf{x}, \mathbf{y}; \theta), \ddot{\mathbf{y}}), \tag{3}$$

$$\mathcal{L}_{intra}^{mt}(\theta) = \text{biKL}(f_1(\mathbf{x}, \mathbf{y}; \theta), f_2(\mathbf{x}, \mathbf{y}; \theta)), \tag{4}$$

$f_1(\cdot)$ and $f_2(\cdot)$ denote the two forward passes of the same model $f(\cdot)$ with the dropout operation, $\text{biKL}(\cdot,\cdot)$ is the bidirectional Kullback-Leibler (KL) divergence of two distributions,

$$\text{biKL}(a, b) = (\text{KL}(a\|b) + \text{KL}(b\|a))/2, \tag{5}$$

$\text{KL}(\cdot\|\cdot)$ denotes the KL divergence of two distributions, and α is a scalar hyper-parameter.

Gao et al. (2023) introduce a cross-lingual consistency regularization, CrossConST, for bridging the representation gap among different languages and improving zero-shot translation in multilingual NMT. For each sentence pair $(\mathbf{x}, \mathbf{y})$, the training objective is defined as:

$$\mathcal{L}_{CrossConST}(\theta) = \mathcal{L}_{ce}^{mt}(\theta) + \beta \mathcal{L}_{cross}^{mt}(\theta), \tag{6}$$

where

$$\mathcal{L}_{cross}^{mt}(\theta) = \text{KL}(f(\mathbf{x}, \mathbf{y}; \theta) \| f(\mathbf{y}, \mathbf{y}; \theta)), \tag{7}$$

and β is a scalar hyper-parameter.

3 Datasets and Baseline Settings

3.1 Dataset Description

We initially consider `en`→`de` translation for empirical study on consistency regularization in Section 4 and then show further experiments for other translation directions in Section 5. The detailed statistics of all datasets are summarized in Table 8.

3.1.1 ST Datasets

We conduct experiments on MuST-C (Di Gangi et al., 2019), which is a multilingual speech translation dataset containing audio recordings with the corresponding transcriptions and translations from English (`en`) to 8 languages: German (`de`), Spanish (`es`), French (`fr`), Italian (`it`), Dutch (`nl`), Portuguese (`pt`), Romanian (`ro`), and Russian (`ru`). We use `dev` and `tst-COMMON` as the validation and test sets respectively.

3.1.2 MT Datasets

We utilize external MT datasets to boost the E2E ST performance. Specifically, we incorporate WMT13 (Bojar et al., 2013) dataset for `en`→`es`, WMT14 (Bojar et al., 2014) dataset for `en`→`fr`, WMT16 (Bojar et al., 2016) datasets for `en`→`de`/`ro`/`ru`, and OPUS100 (Zhang et al., 2020) datasets for `en`→`it`/`nl`/`pt`. Note that we also use `dev` and `tst-COMMON` in the MuST-C dataset as the validation and test sets for the MT tasks.

3.2 Baseline Settings

We adopt a widely used baseline model, W2V2-Transformer (Ye et al., 2021) in our empirical study (Figure 1), which consists of a learnable acoustic feature extractor before two 1-dimensional convolutional layers and the standard Transformer architecture (Vaswani et al., 2017). We use different language tags at the decoder input to distinguish

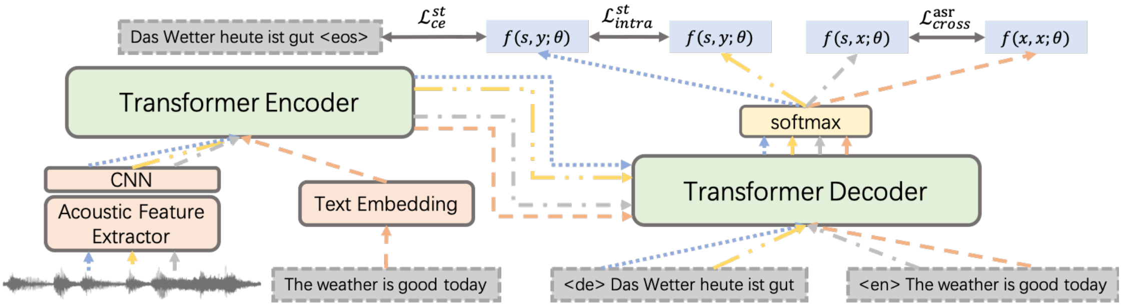

Figure 1: Illustration of the intra-modal and cross-modal consistency regularization. For $\mathcal{L}_{intra}^{st}(\theta)$, the Speech-German pair (Speech, "Das Wetter heute ist gut") goes through the E2E ST model twice and obtain two output distributions $f(\mathbf{s}, \mathbf{y}; \theta)$. For $\mathcal{L}_{cross}^{asr}(\theta)$, the original Speech-English pair (Speech, "The weather is good today") and the copied English-English pair ("The weather is good today", "The weather is good today") go through the E2E ST model and the NMT model respectively and obtain two output distributions $f(\mathbf{s}, \mathbf{x}; \theta)$ and $f(\mathbf{x}, \mathbf{x}; \theta)$.

the target languages. During inference, the language tag serves as the initial token to predict the output text. For example, if the speech input for the sentence “The weather is good today” is in English, to perform ASR, we use <en> as the initial token and decode “The weather is good today”, while to translate into German, we use <de> as the initial token and decode “Das Wetter heute ist gut”.

Pre-processing For speech input, we utilize the raw 16-bit 16kHz mono-channel audio wave. Following common practice, utterances with less than 1000 frames are removed, and utterances with more than 480000 frames are removed in the training set for GPU efficiency. For each translation direction, we jointly learn a unigram SentencePiece (Kudo and Richardson, 2018) model with size 10K on both the source and target sentences and use it to segment sentences into subwords for MT and ST tasks. For the external MT datasets, we filter out parallel sentences which length ratio exceeds 1.5.

Model Configuration We use wav2vec2.0[1] (Baevski et al., 2020) as the acoustic feature extractor, which is pretrained on the audio data from LibriSpeech (Panayotov et al., 2015). Two 1-dimensional convolutional layers are added following the acoustic feature extractor, with kernel size 5, stride size 2, padding 2, and hidden dimension 1024. We utilize 6-layer transformer encoder and 6-layer transformer decoder. Each of the transformer layers comprises 512 hidden units, 8 attention heads, and 2048 feed-forward hidden units.

Training Configuration We apply cross-entropy loss with label smoothing rate 0.1 and set max tokens per batch to be 4096 for the MT task and 2000000 for the ASR and ST tasks. We use the Adam optimizer with Beta $(0.9, 0.98)$, 4000, 8000, and 4000 warmup updates, and inverse square root learning rate scheduler with initial learning rate $1e^{-4}$, $1e^{-3}$, and $1e^{-4}$ for the ASR, MT, and ST tasks respectively. We apply the same configuration in each stage of the training procedure. During inference, we use beam search decoding with a beam size of 8 with length penalty 1.2, 0.6, 1.8, 1.0, 1.0, 1.4, 1.4, and 0.8 for `en`$\rightarrow$`de`, `es`, `fr`, `it`, `nl`, `pt`, `ro`, and `ru`, respectively. We evaluate the MT and ST tasks by case-sensitive sacreBLEU (Post, 2018). We train all models until convergence on 8 NVIDIA Tesla V100 GPUs. For all the experiments below, we select the saved model state with the best validation performance.

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

4 Methodology

In this section, we formally propose SimRegCR and SimZeroCR, the consistency-based strategies for learning E2E ST models in regular (Section 4.1) and zero-shot (Section 4.2) scenarios respectively. We introduce the details of each part below.

4.1 Consistency Regularization for Regular End-to-End Speech Translation

We here investigate the performance of consistency regularization for the regular scenario, where we learn the E2E ST model by utilizing MT and ST datasets. For each training sample, the loss functions include: $\mathcal{L}_{ce}^{mt}(\theta)$, $\mathcal{L}_{intra}^{mt}(\theta)$, $\mathcal{L}_{ce}^{st}(\theta)$,

$$\mathcal{L}_{intra}^{st}(\theta) = \mathrm{biKL}(f_1(\mathbf{s}, \mathbf{y}; \theta), f_2(\mathbf{s}, \mathbf{y}; \theta)), \quad (8)$$

and

$$\mathcal{L}_{cross}^{mt-st}(\theta) = \mathrm{KL}(f(\mathbf{x}, \mathbf{y}; \theta) \| f(\mathbf{s}, \mathbf{y}; \theta)), \quad (9)$$

ID	Training Stage	Loss Function	MT BLEU	ST BLEU
①	MT train from scratch	$\mathcal{L}_{ce}^{mt}$	29.33	-
②	MT train from scratch	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt}$	32.76	-
③	ST train from scratch	$\mathcal{L}_{ce}^{st}$	-	23.49
④	ST train from scratch	$\mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st}$	-	26.77
⑤	ST finetune on ①	$\mathcal{L}_{ce}^{st}$	-	24.38
⑥	ST finetune on ①	$\mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st}$	-	27.35
⑦	ST finetune on ②	$\mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st}$	-	**27.91**
⑧	MT & ST train from scratch	$\mathcal{L}_{ce}^{mt} + \mathcal{L}_{ce}^{st}$	28.54	23.75
⑨	MT & ST finetune on ①	$\mathcal{L}_{ce}^{mt} + \mathcal{L}_{ce}^{st}$	29.73	23.82
⑩	MT & ST finetune on ①	$\mathcal{L}_{ce}^{mt} + \mathcal{L}_{ce}^{st} + \beta\mathcal{L}_{cross}^{mt-st}$	30.66	26.87
⑪	MT & ST finetune on ②	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt} + \mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st}$	32.70	27.48
⑫	MT & ST finetune on ⑪	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt} + \mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st} + \beta\mathcal{L}_{cross}^{mt-st}$	31.00	27.57
⑬	MT train from scratch†	$\mathcal{L}_{ce}^{mt}$	29.61	-
⑭	MT train from scratch†	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt}$	30.02	-
⑮	MT finetune on ⑬	$\mathcal{L}_{ce}^{mt}$	33.59	-
⑯	MT finetune on ⑭	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt}$	34.11	-
⑰	ST finetune on ⑮	$\mathcal{L}_{ce}^{st}$	-	27.33
⑱	ST finetune on ⑮	$\mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st}$	-	28.96
⑲	ST finetune on ⑯	$\mathcal{L}_{ce}^{st} + \alpha\mathcal{L}_{intra}^{st}$	-	**29.23**

Table 1: Case-sensitive detokenized BLEU scores on the MuST-C `en→de` `tst-COMMON` set. † denotes the MT training is performed on the WMT16 dataset, other MT training is performed on the MuST-C dataset. We mark the best ST BLEU scores in two experimental setups in bold. The choices for α and β are summarized in Table 9.

where (1) and (3) are the cross-entropy loss for the ST and MT tasks respectively, (4) and (8) are the intra-modal consistency regularization for the MT and ST tasks respectively, and (9) denotes the cross-modal consistency regularization between the MT and ST tasks, which could also be regarded as the sequence-level knowledge distillation from the MT model to the ST model (Liu et al., 2019).

4.1.1 Experimental Results

We consider two experimental setups: without external MT data (① - ⑫) and with external MT data (⑬ - ⑲), and summarize the experimental results in Table 1. Note that ⑤ and ⑰ correspond to the W2V2-Transformer baselines in the settings of without and with external MT data respectively. By checking model performance under different combinations of loss function and training strategy, we have the following observations: 1) The intra-modal consistency, $\mathcal{L}_{intra}^{mt}$ and $\mathcal{L}_{intra}^{st}$, could boost the MT (① vs ② ; ⑬ vs ⑭) and ST (③ vs ④) performance. 2) The paradigm of pretraining-finetuning could further improve the ST performance (③ vs ⑤ ; ④ vs ⑦). 3) The multi-task learning achieves similar performance compared with the pretraining-finetuning strategy (③ vs ⑧ ; ⑤ vs ⑨). 4) The cross-modal consistency, $\mathcal{L}_{cross}^{mt-st}$, could improve the ST performance (⑨ vs ⑩ ; ⑪ vs ⑫) but still achieve the sub-optimal performance (⑦ vs ⑫).

4.1.2 Does Intra-modal Consistency Implicitly Bridge the Modality Gap?

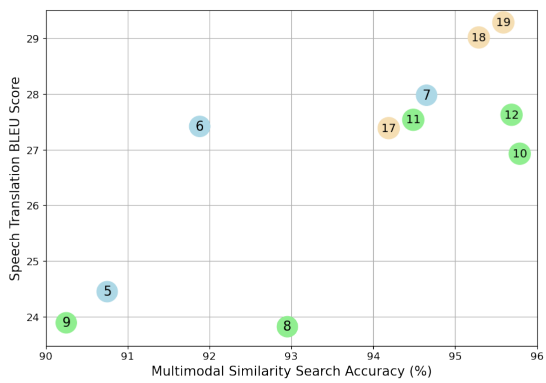

Figure 2: The ST BLEU score and similarity search accuracy of each model in Table 1 on the MuST-C `en→de` `tst-COMMON` set. The blue circles denote the pretraining-finetuning experiments without external MT data. The green circles denote the multi-task learning experiments without external MT data. The orange circles denote the experiments with external MT data.

One interesting finding from the empirical study is that the strategies (⑦ and ⑲) only utilizing the intra-modal consistency achieve the best ST performance instead of explicitly leveraging the cross-modal consistency. We here investigate the impact of the consistency regularization on the modality gap and the E2E ST performance. We conduct a multimodal similarity search experiment and use the averaged bidirectional similarity search accuracy as the metric to evaluate the modality gap. Given parallel speech-transcription pairs, we find the nearest neighbor for each one in the other modality according to the representation cosine similarity and compute the corresponding accuracy, where the speech and transcription representations are calculated by max-pooling the encoder outputs. The evaluation results are reported in Figure 2. By checking the relationship between ST BLEU score and multimodal similarity search accuracy, we have the following observations: 1) The intra-modal consistency, $\mathcal{L}_{intra}^{mt}$ and $\mathcal{L}_{intra}^{st}$, implicitly closes the modality gap (5 vs 6 vs 7 ; 17 vs 18 vs 19). 2) The cross-modal consistency, $\mathcal{L}_{cross}^{mt-st}$, explicitly bridges the modality gap (9 vs 10 ; 11 vs 12). 3) A closer modality gap does not guarantee a better ST performance (6 vs 10 ; 7 vs 12), and the regularization effect introduced by the intra-modal consistency seems to be more crucial for the regular E2E ST task which is in line with Han et al. (2023).

4.1.3 Training Strategy

We here summarize the multi-stage training strategy, SimRegCR (19 in Table 1), consisting of MT pretraining and ST finetuning with the intra-modal consistency regularization in Figure 3. The setting without external MT data only differs by removing the first step of external MT pretraining.

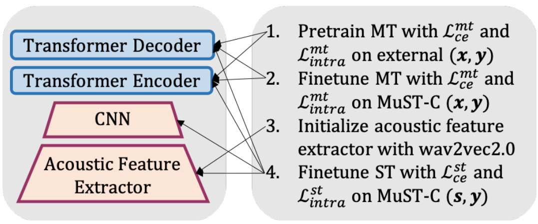

Figure 3: The training steps of SimRegCR by utilizing the intra-modal consistency regularization. In each step, the modules that contribute to the final E2E ST model are pointed out by arrow lines. We also consider SimRegCR$^-$ (18 in Table 1) in this paper, which trains MT model only with $\mathcal{L}_{ce}^{mt}$ in the first two steps.

Comparison with Existing Methods We summarize the recent results of several existing works on the MuST-C en→de benchmark in Table 2. The existing methods vary from different aspects, including cross-modal progressive training (XSTNet) (Ye et al., 2021), cross-modal manifold mixup (STEMM) (Fang et al., 2022), cross-modal contrastive learning (ConST) (Ye et al., 2022), cross-modal mixup via optimal transport (CMOT) (Zhou et al., 2023), and cross-modal regularization with scheduled sampling (CRESS) (Fang and Feng, 2023b). Note that XSTNet, STEMM, and ConST adopt wav2vec2.0 as the acoustic feature extractor, while CMOT and CRESS use HuBERT (Hsu et al., 2021) which could achieve slightly stronger baseline. We can see that SimRegCR$^-$ achieves an improvement of 2.35 BLEU score on average over W2V2-Transformer, and SimRegCR achieves the superior or comparable performance over the current SOTA method CRESS that incorporates cross-modal regularization, scheduled sampling, token-level adaptive training, and a stronger acoustic feature extractor.

Method	BLEU	
	w/o WMT16	w/ WMT16
XSTNet[†]	25.2	27.1
STEMM[†]	25.6	28.7
ConST[†]	25.7	28.3
CMOT[†]	27.0	29.0 / 28.5[*]
CRESS[†]	27.2	**29.4** / 28.9[*]
W2V2-Transformer	24.4	27.3
+ SimRegCR$^-$	27.4	29.0
+ SimRegCR	**27.9**	29.2

Table 2: Our method achieves the superior or comparable performance over the existing methods on the MuST-C en→de benchmark. $*$ denotes the performance of CMOT and CRESS using wav2vec2.0 instead of HuBERT as the acoustic feature extractor. † denotes the numbers are reported from the corresponding papers, others are based on our runs.

4.2 Consistency Regularization for Zero-shot End-to-End Speech Translation

We here investigate the performance of consistency regularization for the zero-shot scenario, where we learn the E2E ST model by utilizing ASR and MT datasets. For each training sample, the loss functions include: $\mathcal{L}_{ce}^{mt}(\theta)$, $\mathcal{L}_{intra}^{mt}(\theta)$,

$$\mathcal{L}_{ce}^{asr}(\theta) = \ell(f(\mathbf{s}, \mathbf{x}; \theta), \ddot{\mathbf{x}}), \quad (10)$$

$$\mathcal{L}_{intra}^{asr}(\theta) = \mathrm{biKL}(f_1(\mathbf{s}, \mathbf{x}; \theta), f_2(\mathbf{s}, \mathbf{x}; \theta)), \quad (11)$$

and

$$\mathcal{L}_{cross}^{asr}(\theta) = \mathrm{KL}(f(\mathbf{s}, \mathbf{x}; \theta) \| f(\mathbf{x}, \mathbf{x}; \theta)), \quad (12)$$

where (3) and (10) are the cross-entropy loss for the MT and ASR tasks respectively, (4) and (11) are the intra-modal consistency regularization for the MT and ASR tasks respectively, and (12) denotes the cross-modal consistency regularization for the ASR task, which could be regarded as the multimodal version of CrossConST (Gao et al., 2023).

ID	Training Stage	Loss Function	MT BLEU	ST BLEU
1	MT train from scratch†	$\mathcal{L}_{ce}^{mt}$	29.61	-
2	MT train from scratch†	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt}$	30.02	-
3	MT Finetune on 1	$\mathcal{L}_{ce}^{mt}$	33.59	-
4	MT Finetune on 2	$\mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt}$	34.11	-
5	ASR & MT finetune on 3	$\mathcal{L}_{ce}^{asr} + \mathcal{L}_{ce}^{mt}$	33.99	0.46
6	ASR & MT finetune on 3	$\mathcal{L}_{ce}^{asr} + \mathcal{L}_{ce}^{mt} + \beta\mathcal{L}_{cross}^{asr}$	32.82	**25.10**
7	ASR & MT finetune on 4	$\mathcal{L}_{ce}^{asr} + \alpha\mathcal{L}_{intra}^{asr} + \mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt}$	34.35	0.56
8	ASR & MT finetune on 7	$\mathcal{L}_{ce}^{asr} + \alpha\mathcal{L}_{intra}^{asr} + \mathcal{L}_{ce}^{mt} + \alpha\mathcal{L}_{intra}^{mt} + \beta\mathcal{L}_{cross}^{asr}$	33.25	24.86

Table 3: Case-sensitive detokenized BLEU scores on the MuST-C `en→de` `tst-COMMON` set. † denotes the MT training is performed on the WMT16 dataset, other MT training is performed on the MuST-C dataset. We mark the best ST BLEU score in bold. The choices for α and β are summarized in Table 10.

4.2.1 Experimental Results

We consider the experimental setup with external MT data and summarize the experimental results in Table 3. Note that 5 corresponds to the W2V2-Transformer baseline. By checking model performance under different combinations of loss function and training strategy, we have the following observations: 1) The cross-modal consistency, $\mathcal{L}_{cross}^{asr}$, could boost the zero-shot ST performance (5 vs 6 ; 7 vs 8). 2) Leveraging the intra-modal consistency, $\mathcal{L}_{intra}^{asr}$ and $\mathcal{L}_{intra}^{mt}$, could improve the corresponding MT performance (5 vs 7 ; 6 vs 8), but could not achieve the superior performance in the zero-shot ST direction (6 vs 8).

4.2.2 Does the Cross-modal Consistency Really Close the Modality Gap?

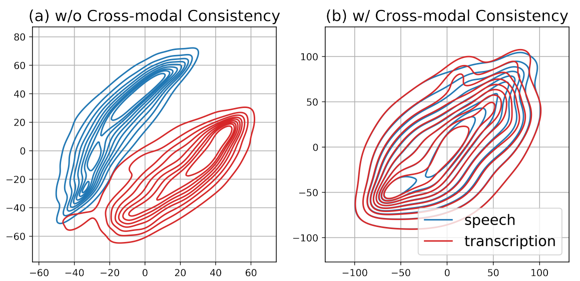

Figure 4: Bivariate kernel density estimation plots of the speech and transcription representations after using T-SNE dimensionality reduction, where the max-pooled outputs of the W2V2-Transformer encoder are applied as the speech and transcription representations.

To verify whether the cross-modal consistency regularization can better align the modality representation space, we visualize the speech and transcription representations of the MuST-C `en→de` `tst-COMMON` set. We apply dimension reduction on the 512-dimensional representations with T-SNE (Hinton and Roweis, 2002) and then depict the bivariate kernel density estimation based on the 2-dimensional representations in Figure 4. Figure 4 shows that the W2V2-Transformer baseline (5) cannot align speech and transcription well in the representation space, while the cross-modal consistency (6) draws the representations across different modalities much closer.

Method	Training Data			BLEU
	Speech	ASR	MT	
MultiSLT†	-	✓	✓	6.8
Chimera†	✓	✓	✓	13.5
DCMA†	✓	✓	✓	24.0
W2V2-Transformer	✓	✓	✓	0.5
+ SimZeroCR	✓	✓	✓	**25.1**

Table 4: Our method achieves the superior performance over the existing methods on the MuST-C `en→de` benchmark. † denotes the numbers are reported from Wang et al. (2022), others are based on our runs.

4.2.3 Training Strategy

We here summarize the multi-stage training strategy, SimZeroCR (6 in Table 3), consisting of MT pretraining and ASR & MT finetuning with the cross-modal consistency regularization in Figure 5.

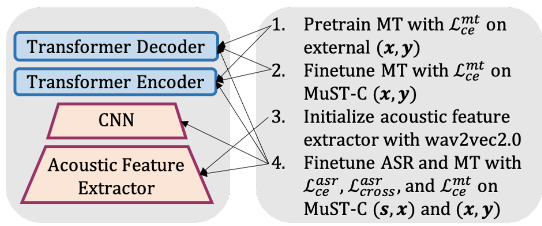

Figure 5: The training steps of SimZeroCR by utilizing the cross-modal consistency regularization. In each step, the modules that contribute to the final E2E ST model are pointed out by arrow lines.

Comparison with Existing Methods We summarize the recent results of several existing works on MuST-C `en→de` benchmark in Table 4. The

Method	External Speech	BLEU							
		de	es	fr	it	nl	pt	ro	ru
Fairseq ST (Wang et al., 2020a)	-	22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3
Dual Decoder (Le et al., 2020)	-	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2
Speechformer (Papi et al., 2021)	-	23.6	28.5	-	-	27.7	-	-	-
SATE (Xu et al., 2021)	-	25.2	-	-	-	-	-	-	-
BiKD (Inaguma et al., 2021)	-	25.3	-	35.3	-	-	-	-	-
XSTNet (Ye et al., 2021)	✓	25.5	29.6	36.0	25.5	30.0	31.3	25.1	16.9
STEMM (Fang et al., 2022)	✓	25.6	30.3	36.1	25.6	30.1	31.0	24.3	17.1
ConST (Ye et al., 2022)	✓	25.7	30.4	36.8	26.3	30.6	32.0	24.8	17.3
FCCL^m (Zhang et al., 2023)	✓	25.9	30.7	36.8	26.4	30.5	31.8	25.0	17.6
M^{3}ST (Cheng et al., 2023)	✓	26.4	31.0	37.2	26.6	30.9	32.8	25.4	18.3
CMOT (Zhou et al., 2023)	✓	27.0	31.1	37.3	26.9	31.2	32.7	25.3	17.9
CRESS (Fang and Feng, 2023b)	✓	27.2	31.9	37.8	27.3	31.6	33.0	25.9	18.7
W2V2-Transformer	✓	24.4	29.9	34.7	25.1	29.3	30.3	23.4	16.5
+ SimRegCR$^-$	✓	27.4	31.5	38.1	27.2	32.0	33.3	25.9	18.8
+ SimRegCR	✓	**27.9***	**32.1***	**39.0***	**27.7***	**32.4***	**34.0***	**26.3***	**19.0***

Table 5: Case-sensitive detokenized BLEU scores on MuST-C tst-COMMON set without external MT datasets. "External speech" denotes unlabeled speech data. * indicates the improvements over W2V2-Transformer are statistically significant with $p < 0.01$. The highest BLEU scores are marked in bold for all methods in each column.

existing methods vary from different aspects, including language-specific encoders-decoders architecture (MultiSLT) (Escolano et al., 2021), continuous cross-modal alignment (Chimera) (Han et al., 2021), and discrete cross-modal alignment (DCMA) (Wang et al., 2022). SimZeroCR achieves an improvement of 24.6 BLEU score over W2V2-Transformer and outperforms the current SOTA method DCMA[2] that incorporates shared memory and vector quantization modules.

5 Experiments on More Languages

5.1 Regular End-to-End Speech Translation

We consider two experimental setups: without external MT data and with external MT data. The detailed information of the baseline methods are summarized in Appendix C, and the BLEU scores of the baseline methods are reported from the corresponding papers. The choice for hyperparameters and the corresponding model performance in each training step of our approaches are summarized in Tables 11, 12, 13, and 14.

When there is no external MT data (Table 5), SimRegCR$^-$ gains an average improvement of 2.6 BLEU scores over the W2V2-Transformer baseline and can achieve comparable performance to the current SOTA method CRESS. It is also worth mentioning that SimRegCR gains an average improvement of 3.1 BLEU scores over the W2V2-Transformer baseline and achieves an average improvement of 0.6 BLEU scores over CRESS that incorporates cross-modal regularization, scheduled sampling, token-level adaptive training, and a stronger acoustic feature extractor, which clearly shows the effectiveness of our methods. When external MT data is included (Table 7), SimRegCR$^-$ and SimRegCR gain average improvement of 1.7 and 2.2 BLEU scores over the W2V2-Transformer baseline respectively, and SimRegCR achieves an average improvement of 0.2 BLEU scores over CRESS, which implies that we could easily achieve SOTA performance for E2E ST task by leveraging simple intra-modal consistency regularization.

5.2 Zero-shot End-to-End Speech Translation

The experimental results with external MT data are summarized in Table 7. For fair comparison, we keep our experimental settings consistent with Wang et al. (2022) to use WMT14 dataset for en→de/es/fr/ru as the external MT data[3]. During inference, we use beam search decoding with a beam size of 5 with length penalty 1.0. The detailed information of the baseline methods are summarized in Appendix D, and the corresponding BLEU scores are reported from Wang et al. (2022). The choice for hyperparameters and the corresponding model performance in each training step of our approach are summarized in Table 15.

Despite the language tag is properly set during inference, W2V2-Transformer is still not capable of translating into specific language and only generating English text. We can see that SimZeroCR gains

[2]Note that the external MT dataset and the inference configurations used in this section are slightly different from those used in Wang et al. (2022). Please check the experimental results in Section 5.2 for more fair comparison.

[3]We only use europarl v7, commoncrawl, and news commentary subsets of WMT14 dataset for en→fr.

Method	External Speech	BLEU de	es	fr	it	nl	pt	ro	ru
MTL (Tang et al., 2021b)	-	23.9	28.6	33.1	-	-	-	-	-
JT-S-MT (Tang et al., 2021a)	-	26.8	31.0	37.4	-	-	-	-	-
Chimera (Han et al., 2021)	✓	27.1	30.6	35.6	25.0	29.2	30.2	24.0	17.4
XSTNet (Ye et al., 2021)	✓	27.1	30.8	38.0	26.4	31.2	32.4	25.7	18.5
STEMM (Fang et al., 2022)	✓	28.7	31.0	37.4	25.8	30.5	31.7	24.5	17.8
ConST (Ye et al., 2022)	✓	28.3	32.0	38.3	27.2	31.7	33.1	25.6	18.9
SpeechUT (Zhang et al., 2022)†	✓	**30.1**	**33.6**	**41.4**	-	-	-	-	-
WACO (Ouyang et al., 2023)	✓	28.1	32.0	38.1	-	-	-	-	-
M^{3}ST (Cheng et al., 2023)	✓	29.3	32.4	38.5	27.5	32.5	33.4	25.9	19.3
FCCL^m (Zhang et al., 2023)	✓	29.0	31.9	38.3	27.3	31.6	32.7	26.8	19.7
CMOT (Zhou et al., 2023)	✓	29.0	32.8	39.5	27.5	32.1	33.5	26.0	19.2
CRESS (Fang and Feng, 2023b)	✓	29.4	33.2	40.1	27.6	32.3	33.6	26.4	19.7
W2V2-Transformer	✓	27.3	31.7	38.0	26.3	29.8	31.7	23.4	18.2
+ SimRegCR$^-$	✓	29.0	33.0	39.4	27.3	32.2	33.5	26.0	19.4
+ SimRegCR	✓	29.2*	33.0*	40.0*	**28.2***	**32.7***	**34.2***	**26.7***	**20.1***

Table 6: Case-sensitive detokenized BLEU scores on MuST-C `tst-COMMON` set with external MT datasets. "External speech" denotes unlabeled speech data. † is a speech-unit-text pretraining model whose training costs are much higher than ours. * indicates the improvements over W2V2-Transformer are statistically significant with $p < 0.01$. The highest BLEU scores are marked in bold for all methods in each column.

Method	BLEU de	es	fr	ru
MultiSLT	6.8	6.8	10.9	-
Chimera	13.5	15.3	22.2	8.3
DCMA	24.0	26.2	33.1	**16.0**
W2V2-Transformer	0.5	0.4	0.4	0.1
+ SimZeroCR	**25.1**	**27.0**	**34.6**	15.6

Table 7: Case-sensitive detokenized BLEU scores on MuST-C `tst-COMMON` set with external MT datasets. The highest BLEU scores are marked in bold for all methods in each column.

an average improvement of 25.2 BLEU scores over the W2V2-Transformer baseline and achieves an average improvement of 0.8 BLEU scores over the current SOTA method DCMA that incorporates shared memory and vector quantization modules, clearly showing the effectiveness of our method.

6 Related Work

E2E ST is a cross-modal task, and one major challenge is direct ST data scarcity. To address such problem, people usually adopt MT data by leveraging the techniques such as pretraining (Bansal et al., 2019; Alinejad and Sarkar, 2020; Le et al., 2021; Tang et al., 2022), multi-task learning (Le et al., 2020; Dong et al., 2021; Indurthi et al., 2021), knowledge distillation (Liu et al., 2019; Gaido et al., 2020; Inaguma et al., 2021), and data augmentation (Lam et al., 2022; Fang and Feng, 2023a). Due to the representation discrepancy between speech and text modalities, people also utilize cross-modal alignment (Han et al., 2021; Fang et al., 2022; Ye et al., 2022; Ouyang et al., 2023) to fully exploit MT data. Specifically, Wang et al. (2022) employ a shared discrete vocabulary space to accommodate both modalities of speech and text and achieve SOTA performance in the zero-shot setting. We show that the zero-shot E2E ST performance could be boosted by leveraging simple cross-modal consistency regularization. Fang and Feng (2023b) propose the cross-modal regularization with scheduled sampling method to bridge the modality gap and achieve the SOTA performance in the regular setting. We find that the regularization is more crucial than modality adaption, which is in line with Han et al. (2023), and achieve the SOTA performance in the regular setting by leveraging simple intra-modal consistency regularization.

7 Conclusion

In this paper, we propose two simple but effective consistency regularization based strategies for learning E2E ST models. We analyze the regularization effect of SimRegCR on the regular E2E ST performance and show that SimZeroCR could effectively close the modality gap. Experiments on the MuST-C benchmark demonstrate the capabilities of our approaches to improve translation performance in both regular and zero-shot settings. Given the universality and simplicity of SimRegCR and SimZeroCR, we believe they can serve as strong baselines for future E2E ST research. For future work, we will explore the effectiveness of consistency regularization on more speech related tasks, such as speech-to-speech translation, speech language modeling, etc.

Limitations

While our approach achieves promising performance by leveraging simple consistency regularization, it still has some limitations: 1) The performance of our approach still lags behind SpeechUT, although the training cost of our approach is much lower. 2) We mainly focus on evaluating our approach on the MuST-C benchmark in this paper. Future research could consider more speech translation benchmarks with more diverse languages, larger ST datasets, and larger models.

References

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020, Online. 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. 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 58–68, Minneapolis, Minnesota. 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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Guandan Chen, Kai Fan, Kaibo Zhang, Boxing Chen, and Zhongqiang Huang. 2021. Manifold adversarial augmentation for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3184–3189, Online. Association for Computational Linguistics.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023. M3st: Mix at three levels for speech translation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. 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.

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

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.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Carlos Segura. 2021. Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 694–701.

Qingkai Fang and Yang Feng. 2023a. Back translation for speech-to-text translation without transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 4567–4587, Toronto, Canada. Association for Computational Linguistics.

Qingkai Fang and Yang Feng. 2023b. Understanding and bridging the modality gap for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15864–15881, Toronto, Canada. Association for Computational Linguistics.

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.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88, Online. 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*, pages 3938–3948, Seattle, United States. Association for Computational Linguistics.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

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.

Yuchen Han, Chen Xu, Tong Xiao, and Jingbo Zhu. 2023. Modality adaption or regularization? a case study on end-to-end speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1340–1348, Toronto, Canada. Association for Computational Linguistics.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

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

Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. 2021. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7723–7727.

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.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824, Online. 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*, volume 34, pages 10890–10905. Curran Associates, Inc.

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 *Proc. Interspeech 2019*, pages 1128–1132.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-aligned contrastive learning for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3891–3907, Toronto, Canada. Association for Computational Linguistics.

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.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Speechformer: Reducing information loss in direct speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1698–1706, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Motoki Sato, Jun Suzuki, and Shun Kiyono. 2019. Effective adversarial regularization for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 204–210, Florence, Italy. Association for Computational Linguistics.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1389, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-attentional models for lattice inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1185–1197, Florence, Italy. Association for Computational Linguistics.

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. 2021a. 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.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Chen Wang, Yuchen Liu, Boxing Chen, Jiajun Zhang, Wei Luo, Zhongqiang Huang, and Chengqing Zong. 2022. Discrete cross-modal alignment enables zero-shot speech translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-End Speech Translation via Cross-Modal Progressive Training. In *Proc. Interspeech 2021*, pages 2267–2271.

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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Wei-Qiang Zhang. 2023. Improving speech translation by cross-modal multi-grained contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1075–1086.

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.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022. 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*, pages 1663–1676, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7873–7887, Toronto, Canada. Association for Computational Linguistics.

Appendix

A Statistics of All Datasets

B The Choice for Hyperparameters in Tables 1 and 3

C Regular E2E ST Methods

We compare our approach with the following methods on the MuST-C benchmark:

- **Fairseq ST** (Wang et al., 2020a): Fairseq ST is a fairseq extension[4] for speech-to-text modeling tasks such as speech translation, which includes end-to-end workflows and SOTA models with scalability and extensibility design.

[4] `https://github.com/facebookresearch/fairseq/tree/main/examples/speech_to_text`

	MuST-C		External MT	
en→	hours	#sents	name	#sents
de	408	234K	WMT16	4.6M
es	504	270K	WMT13	15.2M
fr	492	292K	WMT14	40.8M
it	465	258K	OPUS100	1.0M
nl	442	253K	OPUS100	1.0M
pt	385	211K	OPUS100	1.0M
ro	432	240K	WMT16	0.6M
ru	489	270K	WMT16	2.5M

Table 8: Statistics of all datasets. #sents refers to the number of parallel sentence pairs.

ID	α	β	ID	α	β
1	-	-	2	5.0	-
3	-	-	4	5.0	-
5	-	-	6	5.0	-
7	4.0	-	8	-	-
9	-	-	10	-	5.0
11	3.0	-	12	3.0	5.0
13	-	-	14	0.5	-
15	-	-	16	1.0	-
17	-	-	18	3.0	-
19	3.0	-			

Table 9: The choice for hyperparameters in Table 1.

- **Dual Decoder** (Le et al., 2020): This paper introduces a dual-decoder Transformer architecture for synchronous speech recognition and multilingual speech translation.

- **Speechformer** (Papi et al., 2021): This paper introduces a Transformer-based ST model that able to encode the whole raw audio features without any sub-optimal initial sub-sampling.

- **SATE** (Xu et al., 2021): This paper proposes a stacked acoustic-and-textual encoding method, which is straightforward to incorporate the pre-trained models into ST.

- **BiKD** (Inaguma et al., 2021): To fully leverage knowledge in both source and target language directions for bilingual E2E ST models, this paper proposes bidirectional sequence-level knowledge distillation, in which both forward sequence-level knowledge distillation from a source-to-target

ID	α	β	ID	α	β
①	-	-	②	0.5	-
③	-	-	④	1.0	-
⑤	-	-	⑥	-	45.0
⑦	2.0	-	⑧	2.0	120.0

Table 10: The choice for hyperparameters in Table 3.

NMT model and backward sequence-level knowledge distillation from a target-to-source NMT model are combined.

- **XSTNet** (Ye et al., 2021): This paper proposes cross speech-text network, an extremely concise model which can accept bi-modal inputs and jointly train ST, ASR and MT tasks.
- **MTL** (Tang et al., 2021b): This paper proposes a general multi-task learning framework to leverage text data for ASR and ST tasks.
- **JT-S-MT** (Tang et al., 2021a): This paper proposes three techniques to increase knowledge transfer from the MT task to the ST task, which include parameter sharing and initialization strategy to improve the information sharing between tasks, cross-attentive regularization and online knowledge distillation to encourage the ST system to learn more from the auxiliary MT task and then generate similar model representations from different modalities.
- **STEMM** (Fang et al., 2022): This paper proposes a speech-text manifold mixup method to mix up the speech representation sequences and word embedding sequences.
- **ConST** (Ye et al., 2022): This paper proposes a simple yet effective contrastive learning framework bridging the speech-text representation gap and facilitating the ST with limited data.
- **SpeechUT** (Zhang et al., 2022): This paper proposes a unified-modal speech-unit-text pretraining model, which bridges the modality gap between speech and text representations with hidden units.
- **WACO** (Ouyang et al., 2023): This paper proposes a simple and effective method for extremely low-resource speech-to-text translation, where the key idea is bridging word-level representations for both speech and text modalities via contrastive learning.
- **M^{3}ST** (Cheng et al., 2023): This paper proposes a method to mix the training corpus at three levels, including word level, sentence level and frame level.
- **FCCL^m** (Zhang et al., 2023): This paper proposes a cross-modal multi-grained contrast learning method for explicit knowledge transfer from the MT to the ST model.
- **CMOT** (Zhou et al., 2023): This paper proposes cross-modal mixup via optimal transport to adaptively find the alignment between speech and text sequences, and to mix up the sequences of different modalities at the token level.
- **CRESS** (Fang and Feng, 2023b): This paper proposes a simple yet effective method to regularize the model predictions of ST and MT, whose target-side contexts contain both ground truth words and self-generated words with scheduled sampling.

D Zero-shot E2E ST Methods

We compare our approach with the following methods on the MuST-C benchmark:

- **MultiSLT** (Escolano et al., 2021): This paper extends the multilingual NMT system to perform spoken language translation and zero-shot multilingual spoken language translation by coupling language-specific encoder-decoders, even from monolingual ASR data only.
- **Chimera** (Han et al., 2021): This paper proposes a model capable of learning a text-speech shared semantic memory network for bridging the gap between speech and text representations.
- **DCMA** (Wang et al., 2022): This paper proposes an alignment method to enable zero-shot ST, where the key part is to discretize the continuous vectors to a finite set of virtual tokens and use ASR data to map the corresponding speech and text to the same virtual token in the shared codebook.

E The Choice for Hyperparameters in Section 5

Training Stage		de	es	fr	it	nl	pt	ro	ru
MT pretrain	Baseline	29.33	34.61	41.47	31.25	34.41	35.80	28.13	19.40
ST finetune	Baseline	24.38	29.92	34.73	25.13	29.29	30.32	23.39	16.45
	BLEU	27.35	31.53	38.10	27.24	32.00	33.30	25.89	18.83
	α	5	4	4	5	4	5	4	4

Table 11: The choice for hyperparameters and the corresponding MT & ST performance in the training steps of SimRegCR$^-$ without external MT datasets.

Training Stage		de	es	fr	it	nl	pt	ro	ru
MT pretrain	BLEU	32.76	37.10	45.68	33.31	37.89	39.12	31.60	21.60
	α	5	5	5	5	5	5	5	5
ST finetune	BLEU	27.91	32.12	39.04	27.69	32.39	33.96	26.30	19.02
	α	4	4	5	4	4	4	4	3

Table 12: The choice for hyperparameters and the corresponding MT & ST performance in the training steps of SimRegCR without external MT datasets.

Training Stage		de	es	fr	it	nl	pt	ro	ru
MT pretrain†	Baseline	29.61	31.98	40.59	26.30	30.58	31.83	23.48	18.65
MT finetune	Baseline	33.59	37.78	45.93	32.74	37.06	38.81	29.05	22.11
ST finetune	Baseline	27.33	31.70	38.04	26.29	29.77	31.73	23.43	18.23
	BLEU	28.96	33.04	39.37	27.30	32.22	33.51	26.00	19.41
	α	3	3	2	3	3	4	4	3

Table 13: The choice for hyperparameters and the corresponding MT & ST performance in the training steps of SimRegCR$^-$ with external MT datasets. † denotes the training procedure is performed on the external MT dataset.

Training Stage		de	es	fr	it	nl	pt	ro	ru
MT pretrain†	BLEU	30.02	32.10	40.62	28.24	33.08	34.02	24.99	19.28
	α	0.5	0.25	0.125	3	3	2	2	0.5
MT finetune	BLEU	34.11	37.97	46.95	33.86	38.67	40.09	32.23	22.45
	α	1	0.25	3	5	5	3	3	3
ST finetune	BLEU	29.23	32.97	39.98	28.16	32.68	34.24	26.66	20.09
	α	3	3	3	3	3	4	3	4

Table 14: The choice for hyperparameters and the corresponding MT & ST performance in the training steps of SimRegCR with external MT datasets. † denotes the training procedure is performed on the external MT dataset.

Training Stage		de	es	fr	ru
MT pretrain†	Baseline	29.37	32.91	41.33	18.07
MT finetune	Baseline	33.78	37.53	45.99	21.67
ASR & MT finetune	Baseline	0.47	0.43	0.43	0.07
	BLEU	25.10	26.99	34.59	15.56
	β	30	45	20	35

Table 15: The choice for hyperparameters and the corresponding MT & ST performance in the training steps of SimZeroCR with external MT datasets. † denotes the training procedure is performed on the external MT dataset.