UNDERSTANDING AND BRIDGING THE MODALITY GAP FOR SPEECH TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How to achieve better end-to-end speech translation (ST) by leveraging (text) machine translation (MT) data? Among various existing techniques, multi-task learning is one of the effective ways to share knowledge between ST and MT, thus additional MT data can help to learn the source-to-target mapping. However, due to the differences between speech and text, there is always a gap between ST and MT. In this paper, we first aim to understand this modality gap from the target-side representation differences. We also link the modality gap to another well-known problem in neural machine translation: exposure bias, where the modality gap is relatively small during training except for some hard cases, but keeps increasing during inference due to the cascading effect. To address these problems, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method. Specifically, we regularize the output predictions of ST and MT, whose target-side contexts are derived by sampling between ground truth words and self-generated words with a varying probability. Furthermore, to handle the difficult cases with large modality gaps, we introduce token-level adaptive training to assign different training weights to target tokens according to the extent of the modality gap. Experiments and analysis show that our approach effectively bridges the modality gap, and achieves significant improvements over a strong baseline, which establishes new state-of-the-art results in all eight directions of the MuST-C dataset.[1]

1 INTRODUCTION

End-to-end Speech Translation (ST) aims to translate speech signals to text in another language directly. Compared to traditional cascaded methods, which combine automatic speech recognition (ASR) and machine translation (MT) models in a pipeline manner, end-to-end ST could avoid error propagation and high latency. Recently, end-to-end ST models have achieved comparable or even better results than cascaded ST models (Xu et al., 2021a; Fang et al., 2022).

However, due to the scarcity of ST data, it is difficult to learn a mapping from source speech to target text directly. Previous works often leverage MT data to help the training with multi-task learning (Ye et al., 2022; Tang et al., 2021a). By sharing encoder and decoder between ST and MT, the model tends to learn similar representations from different modalities. In this way, the auxiliary MT task can help build the source-to-target mapping. However, there remains a gap between ST and MT due to the modality gap between speech and text. In this paper, we measure the *modality gap* with the differences between the last decoder layer representations of ST and MT, because the representation of this layer will be mapped into the embedding space to obtain the final translation. A larger modality gap potentially causes different predictions, which makes ST lags behind MT.

Thanks to multi-task learning, we observe that when training with teacher forcing, where both ST and MT use ground truth words as target-side contexts, the modality gap is relatively small except for some difficult cases. However, the exposure bias problem (Bengio et al., 2015; Ranzato et al., 2016) can make things worse. During inference, both ST and MT predict the next token conditioned on their previous generated tokens, which may be different since the modality gap. Moreover, different predictions at the current step may lead to even more different predictions at the next step. As a result, we observe that the modality gap will increase step by step due to this cascading effect.

[1] Code is available in the supplementary material.

To solve these problems, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method. To reduce the effects of exposure bias, we introduce scheduled sampling during training, where the target-side contexts are sampled between ground truth words and self-generated words with a changing probability. Based on this, to bridge the modality gap, we propose to regularize ST and MT in the output space by minimizing a Kullback-Leibler (KL) divergence loss between their predictions. This will encourage greater consistency between ST and MT predictions based on partially self-generated words, which is closer to the inference mode. Besides, to handle the difficult cases, we introduce token-level adaptive training for **CRESS**, where each target token is given a varying weight during training according to the scale of the modality gap. In this way, cases with a significant modality gap will be emphasized. We conduct experiments on the ST benchmark dataset MuST-C (Di Gangi et al., 2019a). Results show that our approach significantly outperforms the strong multi-task learning baseline, with 1.8 BLEU improvements in the base setting and 1.3 BLEU improvements in the expanded setting on average, which establishes new state-of-the-art results in all eight translation directions. Further analysis shows that our approach effectively bridges the modality gap, and improves the translation quality, especially for long sentences.

2 BACKGROUND

2.1 END-TO-END SPEECH TRANSLATION

End-to-end Speech Translation (ST) aims to translate speech in the source language to text in the target language directly. The corpus of ST is usually composed of triplet data $\mathcal{D} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$. Here $\mathbf{s} = (s_1, ..., s_{|\mathbf{s}|})$ is the sequence of audio wave, $\mathbf{x} = (x_1, ..., x_{|\mathbf{x}|})$ is the transcription and $\mathbf{y} = (y_1, ..., y_{|\mathbf{y}|})$ is the translation. Similar to previous work (Ye et al., 2021; Fang et al., 2022), our ST model is composed of an acoustic encoder and a translation model. The acoustic encoder is a pre-trained HuBERT (Hsu et al., 2021) followed by two convolutional layers to reduce the length of the speech sequence. The translation model follows standard Transformer (Vaswani et al., 2017) encoder-decoder architecture, where the encoder contains N Transformer encoder layers, and the decoder contains N Transformer decoder layers. The translation model is first pre-trained with text translation data. The whole model is optimized by minimizing a cross-entropy loss:

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{s}, \mathbf{y}_{<i}), \tag{1}$$

$$p(y_i|\mathbf{s}, \mathbf{y}_{<i}) \propto \exp(\mathbf{W} \cdot f(\mathbf{s}, \mathbf{y}_{<i})), \tag{2}$$

where f is a mapping from the input speech $\mathbf{s}$ and target prefix $\mathbf{y}_{<i}$ to the representation of the last decoder layer at step i. $\mathbf{W}$ is used to transform the dimension to the size of the target vocabulary.

2.2 MULTI-TASK LEARNING FOR SPEECH TRANSLATION

Multi-task learning (MTL) has been proven useful to share knowledge between text translation and speech translation (Tang et al., 2021a), where an auxiliary MT task is introduced during training:

$$\mathcal{L}_{\mathrm{MT}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{x}, \mathbf{y}_{<i}), \tag{3}$$

$$p(y_i|\mathbf{x}, \mathbf{y}_{<i}) \propto \exp(\mathbf{W} \cdot f(\mathbf{x}, \mathbf{y}_{<i})). \tag{4}$$

Note that both modalities (*i.e.,* speech and text) share all transformer encoder and decoder layers. Finally, the training objective is written as follows:

$$\mathcal{L}_{\mathrm{MTL}} = \mathcal{L}_{\mathrm{ST}} + \mathcal{L}_{\mathrm{MT}}. \tag{5}$$

3 PRELIMINARY STUDIES ON THE MODALITY GAP

With multi-task learning, most of the knowledge of MT can be transferred to ST. However, the performance gap between ST and MT still exists. In this section, we first did some preliminary studies with our multi-task learning baseline to understand where this gap comes from.

3.1 Definition of the Modality Gap

The gap between ST and MT is related to the prediction difference at each decoding step, while the prediction depends only on the representation of the last decoder layer. Therefore, we define the *modality gap* at the i-th decoding step as follows:

$$G(\mathbf{s}, \mathbf{y}_{<i}, \mathbf{x}, \mathbf{y}_{<i}) = 1 - cos(f(\mathbf{s}, \mathbf{y}_{<i}), f(\mathbf{x}, \mathbf{y}_{<i})), \tag{6}$$

where cos is the cosine similarity function $cos(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / \|\mathbf{a}\| \|\mathbf{b}\|$. A larger cosine similarity indicates a smaller modality gap.

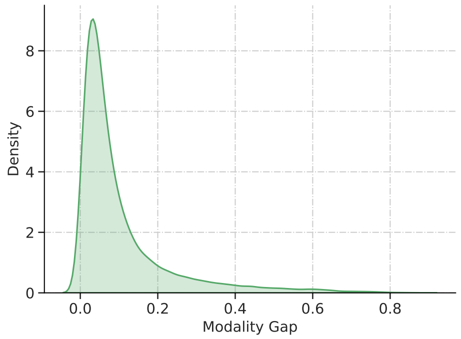

Figure 1: Distribution of the modality gap on MuST-C En→De `dev` set with kernel density estimation (KDE).

To understand the extent of the modality gap, we count the frequency of $G(\mathbf{s}, \mathbf{y}_{<i}, \mathbf{x}, \mathbf{y}_{<i})$ based on all triples $(\mathbf{s}, \mathbf{x}, \mathbf{y}_{<i})$ in MuST-C (Di Gangi et al., 2019a) En→De `dev` set. As shown in Figure 1, the modality gap is relatively small in most cases ($< 10\%$), which proves the effectiveness of multi-task learning in sharing knowledge across ST and MT. However, we also observe a long-tail problem: there is a large difference between ST and MT representations in some difficult cases.

3.2 Connection between Exposure Bias and Modality Gap

Exposure bias, a discrepancy between training and inference, is a well-known problem in neural machine translation (Bengio et al., 2015; Ranzato et al., 2016; Wang & Sennrich, 2020; Arora et al., 2022). During training with *teacher forcing*, both ST and MT predict the next token conditioned on the ground truth target prefix $\mathbf{y}_{<i}$. However, during inference, the predictions of ST and MT depend on their previous generated tokens by the model itself (denoted as $\widehat{\mathbf{y}}^s_{<i}$ and $\widehat{\mathbf{y}}^x_{<i}$ for ST and MT respectively), which might be different due to the modality gap. Furthermore, different predictions at the current decoding step result in different target prefixes for ST and MT, potentially causing even more different predictions at the next step. The such cascading effect will enlarge the modality gap step by step during inference.

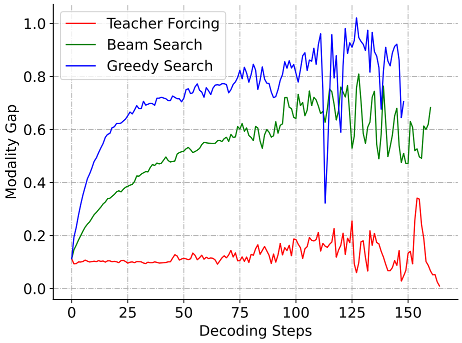

Figure 2: Curves of the average modality gap on MuST-C En→De `dev` set with decoding steps under *teacher forcing*, *beam search*, and *greedy search* strategies. Note that for *beam search* we have several candidate translations. The modality gap is calculated with the average representation of all candidates. Here we set a beam size of 8.

To prove our hypothesis, we present the curves of the modality gap with decoding steps under *teacher forcing*, *beam search*, and *greedy search* strategies, respectively. As shown in Figure 2, with *teacher forcing*, there is no significant difference in the modality gap across steps, as both ST and MT depend on the same target prefix at any step, so the modality gap $G(\mathbf{s}, \mathbf{y}_{<i}, \mathbf{x}, \mathbf{y}_{<i})$ only comes from the difference between input speech $\mathbf{s}$ and text $\mathbf{x}$. However, when decoding with *greedy search*, due to the cascading effect mentioned above, the self-generated target prefix $\widehat{\mathbf{y}}^s_{<i}$ and $\widehat{\mathbf{y}}^x_{<i}$ become more and more different, which makes the modality gap $G(\mathbf{s}, \widehat{\mathbf{y}}^s_{<i}, \mathbf{x}, \widehat{\mathbf{y}}^x_{<i})$ keep increasing with decoding steps. A simple way to alleviate this problem is *beam search*, which considers several candidate tokens rather than a single one at each decoding step. When there is an overlap between candidate tokens of ST and MT, the cascading effect will be reduced, thus slowing down the increase of the modality gap.

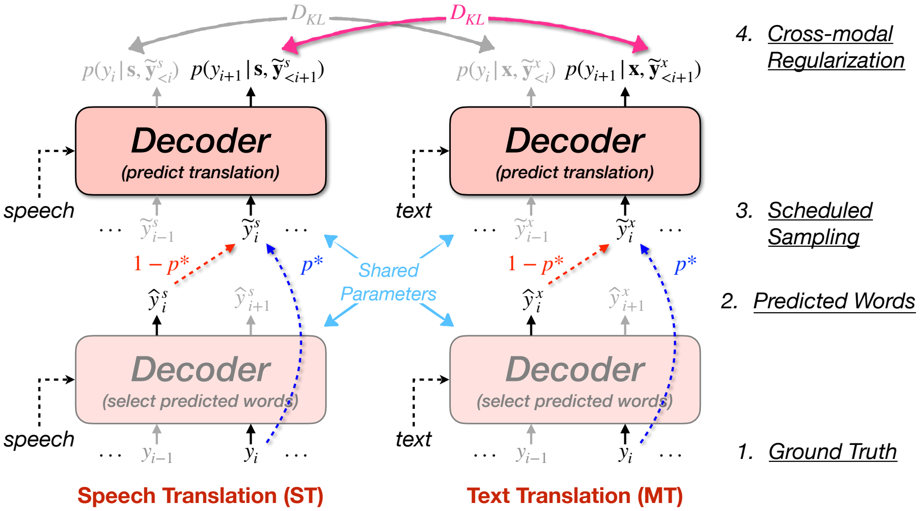
D_{KL}
D_{KL}
$p(y_i \mid \mathbf{s}, \widetilde{\mathbf{y}}^s_{<i})$ $p(y_{i+1} \mid \mathbf{s}, \widetilde{\mathbf{y}}^s_{<i+1})$
$p(y_i \mid \mathbf{x}, \widetilde{\mathbf{y}}^x_{<i})$ $p(y_{i+1} \mid \mathbf{x}, \widetilde{\mathbf{y}}^x_{<i+1})$
4. Cross-modal Regularization
Decoder
(predict translation)
Decoder
(predict translation)
speech
text
$\dots$ $\widetilde{y}^s_{i-1}$ $\widetilde{y}^s_i$ $\dots$
$\dots$ $\widetilde{y}^x_{i-1}$ $\widetilde{y}^x_i$ $\dots$
3. Scheduled Sampling
$1-p^*$
p^*
Shared Parameters
$1-p^*$
p^*
$\widehat{y}^s_i$ $\widehat{y}^s_{i+1}$
$\widehat{y}^x_i$ $\widehat{y}^x_{i+1}$
2. Predicted Words
Decoder
(select predicted words)
Decoder
(select predicted words)
speech
text
$\dots$ y_{i-1} y_i $\dots$
$\dots$ y_{i-1} y_i $\dots$
1. Ground Truth
Speech Translation (ST)
Text Translation (MT)

Figure 3: Overview of our proposed **CRESS**. Note that the step of selecting predicted words has no gradient calculation and is fully parallelized.

4 METHOD: CRESS

Our preliminary studies in Section 3 show that:

- The modality gap will be enlarged during inference due to exposure bias.
- The modality gap may be significant in some difficult cases.

Inspired by these, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method, to bridge the modality gap, especially in inference mode (Section 4.1). Furthermore, we propose a token-level adaptive training method for **CRESS** to handle the difficult cases (Section 4.2).

4.1 CROSS-MODAL REGULARIZATION WITH SCHEDULED SAMPLING

To bridge the modality gap during inference, we adopt scheduled sampling for both ST and MT to approximate the inference mode at training time. After that, we add a regularization loss between the predictions of ST and MT based on part of their self-generated words as context. This allows for more consistent predictions between ST and MT during inference, thus reducing the performance gap between ST and MT. Figure 3 illustrates the main framework of our method.

Scheduled Sampling *Scheduled sampling* (Bengio et al., 2015), which samples between ground truth words and self-generated words, i.e. *predicted words*, with a certain probability as target-side context, has proven useful for alleviating exposure bias. In general, the input at the $\{i+1\}$-th decoding step should be the ground truth word y_i during training. With scheduled sampling, it can also be substituted by a predicted word. Next we describe how to select the predicted word $\widehat{y}^s_i$ for ST and $\widehat{y}^x_i$ for MT. For ST, we follow Zhang et al. (2019) to select the predicted word $\widehat{y}^s_i$ by sampling from the word distribution $p(y_i|\mathbf{s}, \mathbf{y}_{<i})$ in Equation (2) with *Gumbel-Max* technique (Gumbel, 1954; Maddison et al., 2014), a method to draw a sample from a categorical distribution:

$$\eta = -\log(-\log u), \tag{7}$$
$$\widehat{y}^s_i = \arg\max\left(\mathbf{W} \cdot f(\mathbf{s}, \mathbf{y}_{<i}) + \eta\right), \tag{8}$$

where η is the Gumbel noise calculated from the uniform noise $u \sim \mathcal{U}(0,1)$. Similarly, for MT, there is:

$$\widehat{y}^x_i = \arg\max\left(\mathbf{W} \cdot f(\mathbf{x}, \mathbf{y}_{<i}) + \eta\right). \tag{9}$$

Note that we may omit the superscript and denote the predicted word for both ST and MT by $\widehat{y}_i$ in the following.

How to select between the ground truth word y_i and the predicted word $\widehat{y}_i$? Similar to Bengio et al. (2015); Zhang et al. (2019), we randomly sample from both of them with a varying probability. We denote the probability of selecting from the ground truth word as p^*. At the beginning of training, since the model is not yet well trained, we select more from the ground truth words (with larger p^*) to help the model converge. In the later stages of training, we select more from the predicted words (with smaller p^*), which is closer to the situation during inference. To achieve this, we decrease p^* with a function of the index of training epochs e:

$$p^* = \frac{\mu}{\mu + \exp(e/\mu)}, \tag{10}$$

where μ is a hyper-parameter. With scheduled sampling, the target-side context becomes $\widetilde{\mathbf{y}} = (\widetilde{y}_1, ..., \widetilde{y}_{|\mathbf{y}|})$, where

$$\widetilde{y}_i = \begin{cases} y_i, & p \leq p^* \\ \widehat{y}_i, & p > p^* \end{cases}, \tag{11}$$

where p is sampled from the uniform distribution $\mathcal{U}(0, 1)$. Using $\widetilde{\mathbf{y}}^s$ and $\widetilde{\mathbf{y}}^x$ to denote the target-side context of ST and MT respectively, the loss functions of ST and MT become:

$$\mathcal{L}_{\mathrm{ST}}^{\textsc{Cress}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}), \tag{12}$$

$$\mathcal{L}_{\mathrm{MT}}^{\textsc{Cress}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i}), \tag{13}$$

Cross-modal Regularization To bridge the modality gap in inference mode, we expect the predictions of ST and MT with scheduled sampling to be consistent. Inspired by recent works of consistency training (liang et al., 2021; Guo et al., 2022), we regularize ST and MT in the output space. Specifically, we minimize the bidirectional Kullback-Leibler (KL) divergence between the output distributions of ST and MT at each decoding step:

$$\mathcal{L}_{\mathrm{Reg}}^{\textsc{Cress}} = \sum_{i=1}^{|\mathbf{y}|} \frac{1}{2}(\mathcal{D}_{\mathrm{KL}}(p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}) \| p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i})) + \mathcal{D}_{\mathrm{KL}}(p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i}) \| p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}))). \tag{14}$$

With the translation loss in Equation (12) and (13), the final training objective is as follows:

$$\mathcal{L}^{\textsc{Cress}} = \mathcal{L}_{\mathrm{ST}}^{\textsc{Cress}} + \mathcal{L}_{\mathrm{MT}}^{\textsc{Cress}} + \lambda \mathcal{L}_{\mathrm{Reg}}^{\textsc{Cress}}, \tag{15}$$

where λ is the hyper-parameter to control the weight of $\mathcal{L}_{\mathrm{Reg}}^{\textsc{Cress}}$.

4.2 Token-level Adaptive Training for **Cress**

As we mentioned above, the modality gap might be significant in some difficult cases. Inspired by the idea of token-level adaptive training (Gu et al., 2020; Xu et al., 2021b; Zhang et al., 2022b), we propose to treat each token adaptively according to the scale of the modality gap. The training objectives in Equation (12), (13), and (14) are modified as follows:

$$\mathcal{L}_{\mathrm{ST}}^{\textsc{Cress}} = -\sum_{i=1}^{|\mathbf{y}|} w_i \cdot \log p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}), \tag{16}$$

$$\mathcal{L}_{\mathrm{MT}}^{\textsc{Cress}} = -\sum_{i=1}^{|\mathbf{y}|} w_i \cdot \log p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i}), \tag{17}$$

$$\mathcal{L}_{\mathrm{Reg}}^{\textsc{Cress}} = \sum_{i=1}^{|\mathbf{y}|} \frac{1}{2} w_i \cdot (\mathcal{D}_{\mathrm{KL}}(p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}) \| p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i})) + \mathcal{D}_{\mathrm{KL}}(p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i}) \| p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}))), \tag{18}$$

where w_i is the token-level weight defined by a linear function of the modality gap:

$$w_i = B + S \cdot G(\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}, \mathbf{x}, \widetilde{\mathbf{y}}^x_{<i}), \tag{19}$$

where B (base) and S (scale) are two hyper-parameters to control the lower bound and magnitude of change of w_i. In this way, cases with a large modality gap will be assigned a larger weight and thus emphasized during training. Note that the modality gap is computed on-the-fly during training.

5 Experiments

5.1 Datasets

ST Datasets We conduct experiments on MuST-C (Di Gangi et al., 2019a) dataset, a multilingual speech translation dataset containing 8 translation directions: English (En) to German (De), French (Fr), Spanish (Es), Romanian (Ro), Russian (Ru), Italian (It), Portuguese (Pt) and Dutch (Nl). It contains at least 385 hours of TED talks with transcriptions and translations for each direction. We use `dev` set for validation and `tst-COMMON` set for evaluation.

External MT Datasets We also introduce external MT datasets to pre-train our translation model in the expanded setting. For En→De/Fr/Es/Ro/Ru directions, we introduce data from WMT. For En→It/Pt/Nl, we introduce data from OPUS100[2] (Zhang et al., 2020).

Tabel 3 in Appendix A lists the statistics of all ST and external MT datasets.

5.2 Experimental Setups

Pre-processing For *speech* input, we use the raw 16-bit 16kHz mono-channel audio wave. For *text* input, all sentences in ST and external MT datasets are tokenized and segmented into subwords using SentencePiece[3]. For each translation direction, the vocabulary is learned from the source and target texts from the ST dataset, with a size of 10K. For the external MT datasets (WMT and OPUS100), we filter out parallel sentence pairs whose length ratio exceeds 1.5.

Model Setting We use the pre-trained HuBERT model[4] to encode the input audio. Two 1-dimensional convolutional layers after HuBERT are set to kernel size 5, stride size 2, and padding 2. For the translation model, we employ post-LN (layer normalization) Transformer architecture with the base configuration, which contains 6 encoder layers and 6 decoder layers, with 512 hidden states, 8 attention heads, and 2048 feed-forward hidden states for each layer. The translation model is first pre-trained with MT task using *transcription-translation* pairs from the ST dataset (**base setting**), and also sentence pairs from the external MT dataset (**expanded setting**).

During MT pre-training, each batch has up to 33k text tokens. The maximum learning rate is set to 7e-4. During fine-tuning, each batch contains up to 16M audio frames. The maximum learning rate is set to 1e-4. We use Adam optimizer (Kingma & Ba, 2015) with 4k warm-up steps. We set dropout to 0.1 and label smoothing to 0.1. During inference, we average the checkpoints of the last 10 epochs for evaluation. We use beam search with a beam size of 8. We use scareBLEU[5] (Post, 2018) to compute case-sensitive detokenized BLEU (Papineni et al., 2002) scores and the statistical significance of translation results with paired bootstrap resampling[6] (Koehn, 2004). We implement our model with *fairseq*[7] (Ott et al., 2019). All models are trained on 4 Nvidia RTX 3090 GPUs.

For scheduled sampling, the decay parameter is $\mu = 15$ (See Appendix B for details). For cross-modal regularization, the weight parameter is $\lambda = 1.0$. For token-level adaptive training, we did a grid search for base and scale parameters on MuST-C En→De `dev` set with $B \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ and $S \in \{0.05, 0.10, 0.20, 0.50, 1.00\}$. Finally, we set $B = 0.7$ and $S = 0.05$ for all translation directions.

Baseline Systems We include several strong end-to-end ST systems for comparison: Fairseq ST (Wang et al., 2020a), RevisitST (Zhang et al., 2022a), DDT (Le et al., 2020), LAT (Le et al., 2021), Chimera (Han et al., 2021), XSTNet (Ye et al., 2021), TDA (Du et al., 2022), STEMM (Fang et al., 2022), ConST (Ye et al., 2022), TaskAware (Indurthi et al., 2021), STPT (Tang et al., 2022). Besides, the multi-task learning baseline in Section 2.2 is also included as a strong baseline, which is denoted as **MTL**. We use **Cress** to denote our method with token-level adaptive training (Section 4.2).

[2] `http://opus.nlpl.eu/opus-100.php`

[3] `https://github.com/google/sentencepiece`

[4] `https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt`

[5] `https://github.com/mjpost/sacrebleu`

[6] sacreBLEU signature:nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

[7] `https://github.com/pytorch/fairseq`

Table 1: BLEU scores on MuST-C `tst-COMMON` set. The external MT datasets are only used in the expanded setting. Scores with grey background indicate the previous state-of-the-art results of each translation direction. * and ** mean the improvements over **MTL** baseline is statistically significant ($p < 0.05$ and $p < 0.01$, respectively). † uses a pre-trained acoustic encoder (*e.g.*, Wav2vec 2.0 or HuBERT). ‡ STPT (Tang et al., 2022) jointly pre-trains speech and text on large-scale datasets.

Models	**BLEU**								
	En→De	En→Fr	En→Es	En→Ro	En→Ru	En→It	En→Pt	En→Nl	Avg.
Base setting (w/o external MT data)									
Fairseq ST (Wang et al., 2020a)	22.7	32.9	27.2	21.9	15.3	22.7	28.1	27.3	24.8
RevisitST (Zhang et al., 2022a)	23.0	33.5	28.0	23.0	15.6	23.5	28.2	27.1	25.2
DDT (Le et al., 2020)	23.6	33.5	28.1	22.9	15.2	24.2	30.0	27.6	25.6
TDA (Du et al., 2022)	25.4	36.1	29.6	23.9	16.4	25.1	31.1	29.6	27.2
†XSTNet (Ye et al., 2021)	25.5	36.0	29.6	25.1	16.9	25.5	31.3	30.0	27.5
†STEMM (Fang et al., 2022)	25.6	36.1	30.3	24.3	17.1	25.6	31.0	30.1	27.5
†ConST (Ye et al., 2022)	25.7	36.8	30.4	24.8	17.3	26.3	32.0	30.6	28.0
†**MTL**	25.3	35.7	30.5	23.8	17.2	26.0	31.3	29.5	27.4
†**CRESS**	**27.2****	**37.8****	**31.9****	**25.9****	**18.7****	**27.3****	**33.0****	**31.6****	**29.2**
Expanded setting (w/ external MT data)									
LAT (Le et al., 2021)	24.7	35.0	28.7	23.8	16.4	25.0	31.1	28.8	26.7
TaskAware (Indurthi et al., 2021)	28.9	-	-	-	-	-	-	-	-
†Chimera (Han et al., 2021)	27.1	35.6	30.6	24.0	17.4	25.0	30.2	29.2	27.4
†XSTNet (Ye et al., 2021)	27.1	38.0	30.8	25.7	18.5	26.4	32.4	31.2	28.8
†STEMM (Fang et al., 2022)	28.7	37.4	31.0	24.5	17.8	25.8	31.7	30.5	28.4
†ConST (Ye et al., 2022)	28.3	38.3	32.0	25.6	18.9	27.2	33.1	31.7	29.4
‡STPT (Tang et al., 2022)	-	39.7	33.1	-	-	-	-	-	-
†**MTL**	27.7	38.5	32.8	24.9	19.0	26.5	32.0	30.8	29.0
†**CRESS**	**29.4****	**40.1****	**33.2***	**26.4****	**19.7****	**27.6****	**33.6****	**32.3****	**30.3**

5.3 MAIN RESULTS ON MUST-C DATASET

Table 1 shows the results on MuST-C `tst-COMMON` set in all eight directions. First, we noticed that our implemented **MTL** is a relatively strong baseline compared with existing approaches. Second, our proposed **CRESS** significantly outperforms **MTL** in both settings, with 1.8 BLEU improvement in the base setting and 1.3 BLEU improvement in the expanded setting on average, which establishes new state-of-the-art results on MuST-C dataset, demonstrating the superiority of our approach. Besides, we also report ChrF++ scores in Appendix E.

6 ANALYSIS AND DISCUSSION

Results in Section 5.3 show the superiority of our method. To better understand **CRESS**, we explore several questions in this section. All analysis experiments are conducted on MuST-C En→De `dev` set in the expanded setting.

(1) Do scheduled sampling, cross-modal regularization, and token-level adaptive training all matter? Scheduled sampling, regularization, and token-level adaptive training are effective techniques to improve the performance of translation models. To understand the role of each, we conduct ablation experiments in Table 2. When training with scheduled sampling only (Line 4), we observe a slight improvement of 0.1 BLEU, which is probably due to the alleviation of exposure bias. When training with cross-modal regularization only (Line 3), which encourages the consistency between predictions of ST and MT with ground truth target side context, we observe an improvement of 0.5 BLEU. If we combine both together (Line 2), we obtain a much more significant boost of 1.1 BLEU,

Table 2: BLEU scores on MuST-C En→De `dev` set with different combinations of training techniques. **TAT.** indicates token-level adaptive training. **Reg.** indicates cross-modal regularization. **SS.** indicates scheduled sampling. * and ** mean the improvements over **MTL** baseline (Line 5) is statistically significant ($p < 0.05$ and $p < 0.01$, respectively).

TAT.	**Reg.**	**SS.**	**BLEU**
✓	✓	✓	**28.4****
×	✓	✓	28.0**
×	✓	×	27.4*
×	×	✓	27.0
×	×	×	26.9

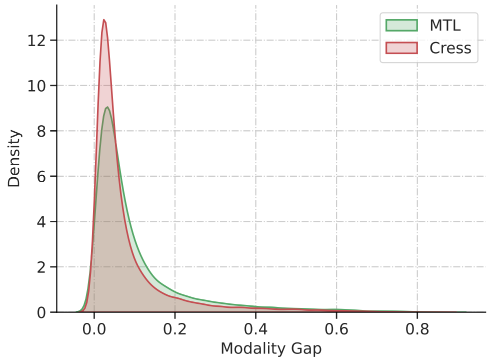

Figure 4: Distributions of the modality gap on MuST-C En→De `dev` set of **MTL** and **CRESS** with kernel density estimation (KDE).

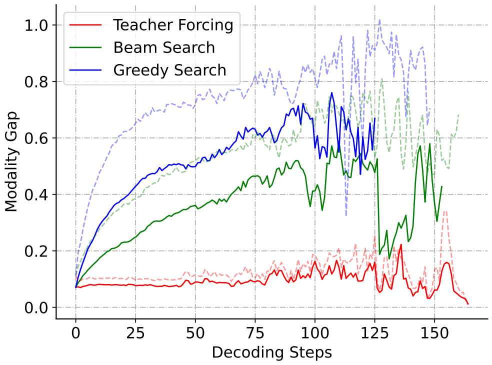

Figure 5: Curves of the average modality gap with decoding steps under three strategies. The dotted line refers to **MTL** (same as Figure 2), and the solid line refers to **CRESS**.

which proves that both scheduled sampling and cross-modal regularization play a crucial role in our method. Furthermore, with token-level adaptive training, the improvement comes to 1.5 BLEU, which shows the benefit of treating different tokens differently according to the modality gap.

(2) Does CRESS successfully bridge the modality gap? To validate whether our approach successfully bridges the modality gap between ST and MT, we revisit the experiments in Section 3. Figure 4 shows the distribution of the modality gap with teacher forcing. We observe a general decrease in the modality gap compared with **MTL**. We also plot the curves of the modality gap with decoding steps of **CRESS** under teacher forcing, greedy search, and beam search strategies. As shown in Figure 5, our approach significantly slows down the increase of the modality gap compared with **MTL** baseline, suggesting that the predictions of ST and MT are more consistent during inference, demonstrating the effectiveness of our method in bridging the modality gap.

(3) How base and scale hyper-parameters influence token-level adaptive training? B (base) and S (scale) are two important hyper-parameters in token-level adaptive training. We investigate how different combinations of B and S influence performance. As shown in Figure 6, token-level adaptive training can bring improvements over **CRESS** in most cases. In particular, it usually performs better with smaller B and smaller S, leading to a boost of up to 0.4 BLEU. We conclude that treating different tokens too differently is also undesirable.

(4) Is CRESS more effective for longer sentences? The autoregressive model generates the translation step by step, so the translation of long sentences would be more challenging. We divide the MuST-C En→De `dev` set into several groups according to the length of target sentences, and compute the BLEU scores in each group separately, which is shown in Figure 7. We observe that **CRESS** achieve large improvements over the baseline in all groups, especially for sentences longer than 45, which shows the superiority of our method when translating long sentences.

7 RELATED WORK

End-to-end Speech Translation End-to-end speech translation (Bérard et al., 2016; Weiss et al., 2017) has shown great potential for overcoming error propagation and reducing latency compared to traditional cascaded ST systems (Salesky et al., 2019; Di Gangi et al., 2019b;c; Bahar et al., 2019a). One challenge in training end-to-end ST models is the scarcity of ST data. To address this problem, researchers employed MT data to help training with techniques like pre-training (Bansal et al., 2019; Stoian et al., 2020; Wang et al., 2020c;d; Alinejad & Sarkar, 2020; Le et al., 2021; Dong et al., 2021a; Zheng et al., 2021; Xu et al., 2021a; Tang et al., 2022), multi-task learning (Le et al., 2020; Dong et al., 2021b; Ye et al., 2021; Tang et al., 2021a;b; Indurthi et al., 2021), knowledge distillation (Liu et al., 2019; Inaguma et al., 2021), and data augmentation (Jia et al., 2019; Bahar et al., 2019b; Lam et al., 2022). However, due to the *modality gap* between speech and text, it is still difficult to fully exploit MT data with the above techniques. To overcome the modality gap,

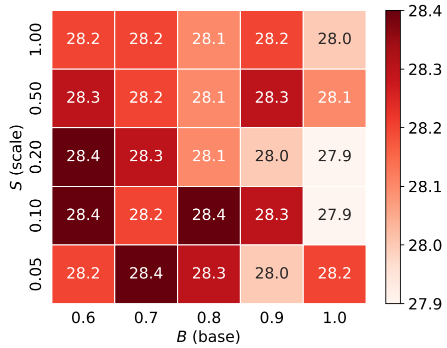

Figure 6: The heat map of BLEU scores on MuST-C En→De `dev` set with different combinations of B and S. The BLEU score without token-level adaptive training is 28.0.

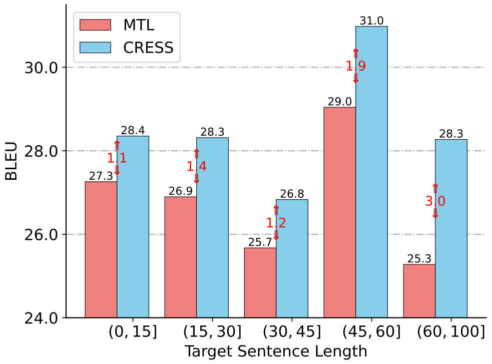

Figure 7: BLEU scores on MuST-C En→De `dev` set at different target sentence lengths.

Han et al. (2021) projects features of both speech and text into a shared semantic space. Fang et al. (2022) introduces mixup between features of speech and text to learn similar representations for them. Ye et al. (2022) brings sentence-level representations closer with cross-modal contrastive learning. Bapna et al. (2021; 2022); Chen et al. (2022); Tang et al. (2022) jointly train on speech and text and design training objectives to align two modalities. Different from previous work, in this work, we understand the modality gap from the target-side representation differences, and show its connection to exposure bias. Based on this, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method to bridge the modality gap during inference.

Exposure Bias Exposure bias indicates the discrepancy between training and inference. Several approaches employ Reinforcement Learning (RL) (Ranzato et al., 2016; Shen et al., 2016; Bahdanau et al., 2017) instead of Maximum Likelihood Estimation (MLE) to avoid this problem. However, Wu et al. (2018) shows that RL-based training is unstable due to the high variance of gradient estimation. An alternative and simpler approach is scheduled sampling (Bengio et al., 2015), which samples between ground truth words and self-generated words with a changing probability. Zhang et al. (2019) extends it with Gumbel noise for more robust training. In this paper, we adopt this approach to approximate the inference mode due to its training stability and low training cost.

Output Regularization for MT Regularization on the output space has proved useful for machine translation. liang et al. (2021) proposes to regularize the output predictions of two sub-models sampled by dropout. Guo et al. (2022) regularizes the output predictions of models before and after input perturbation. In this paper, we turn to regularize the output predictions across modalities, which encourages more consistent predictions for ST and MT.

Token-level Adaptive Training Token-level adaptive training for MT is first proposed in Gu et al. (2020), which assigns larger weights to low-frequency words to prevent them from being ignored. Xu et al. (2021b); Zhang et al. (2022b) computes the weight with bilingual mutual information. In this paper, we compute the weights with the modality gap between ST and MT.

8 CONCLUSION

In this paper, we propose a simple yet effective method **CRESS** 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. Based on this, we further propose a token-level adaptive training method to handle difficult cases. Our method establishes new state-of-the-art results on MuST-C dataset. Further analysis shows that our method can effectively bridge the modality gap and improve the translation quality, especially for long sentences. In the future, we will explore how to apply our method to other tasks.

REFERENCES

Ashkan Alinejad and Anoop Sarkar. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8014–8020, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.644. URL https://aclanthology.org/2020.emnlp-main.644.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 4218–4222. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 700–710, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.58. URL https://aclanthology.org/2022.findings-acl.58.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: self-supervised cross-lingual speech representation learning at scale. In Hanseok Ko and John H. L. Hansen (eds.), *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pp. 2278–2282. ISCA, 2022. doi: 10.21437/Interspeech.2022-143. URL https://doi.org/10.21437/Interspeech.2022-143.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 2020.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. A comparative study on end-to-end speech to text translation. In *Proc. of ASRU*, pp. 792–799. IEEE, 2019a.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. On using specaugment for end-to-end speech translation. In *Proc. of IWSLT*, 2019b.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJDaqqveg.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proc. of NAACL-HLT*, pp. 58–68, 2019.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*, 2021.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. mslam: Massively multilingual joint pre-training for speech and text. *CoRR*, abs/2202.01374, 2022. URL https://arxiv.org/abs/2202.01374.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. Maestro: Matched speech text representations through modality matching. In *INTERSPEECH*, 2022.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 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)*, pp. 2012–2017, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://aclanthology.org/N19-1202.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. Adapting transformer to end-to-end spoken language translation. In *Proc. of INTERSPEECH*, pp. 1133–1137. International Speech Communication Association (ISCA), 2019b.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. Enhancing transformer for end-to-end speech-to-text translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pp. 21–31, 2019c.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. Consecutive decoding for speech-to-text translation. In *The Thirty-fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021a.

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

Yichao Du, Zhirui Zhang, Weizhi Wang, Boxing Chen, Jun Xie, and Tong Xu. Regularizing end-to-end speech translation with triangular decomposition agreement. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 10590–10598. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21303.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 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)*, pp. 7050–7062, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.486. URL https://aclanthology.org/2022.acl-long.486.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In *4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014*, pp. 16–23. ISCA, 2014. URL http://www.isca-speech.org/archive/sltu_2014/gales14_sltu.html.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1035–1046, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.76. URL https://aclanthology.org/2020.emnlp-main.76.

Emil Julius Gumbel. Statistical theory of extreme values and some practical applications: a series of lectures. In *Nat. Bur. Standards Appl. Math. Ser.*, volume 33. US Government Printing Office, 1954. URL https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB175818.xhtml#.

Dengji Guo, Zhengrui Ma, Min Zhang, and Yang Feng. Prediction difference regularization against perturbation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2214–2225, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.195. URL https://aclanthology.org/2021.findings-acl.195.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, jan 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of NAACL*, pp. 1872–1881, 2021.

Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7723–7727, 2021. doi: 10.1109/ICASSP39728.2021.9414703.

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

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020. https://github.com/facebookresearch/libri-light.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 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)*, pp. 245–254, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.27. URL https://aclanthology.org/2022.acl-short.27.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3520–3533, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.314. URL https://aclanthology.org/2020.coling-main.314.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 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)*, pp. 817–824, Online, August 2021.

Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.103. URL https://aclanthology.org/2021.acl-short.103.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 827–838, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.68. URL https://aclanthology.org/2021.acl-long.68.

xiaobo liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-drop: Regularized dropout for neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/5a66b9200f29ac3fa0ae244cc2a51b39-Paper.pdf.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pp. 1128–1132, 2019. doi: 10.21437/Interspeech.2019-2582.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A$*$ sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/309fee4e541e51de2e41f21bebb342aa-Paper.pdf.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL https://aclanthology.org/W17-4770.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://www.aclweb.org/anthology/W18-6319.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 2757–2761. ISCA, 2020. doi: 10.21437/Interspeech.2020-2826. URL https://doi.org/10.21437/Interspeech.2020-2826.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International*

Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL http://arxiv.org/abs/1511.06732.

Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. Fluent translations from disfluent speech in end-to-end speech translation. In *Proc. of NAACL-HLT*, pp. 2786–2792, 2019.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1683–1692, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1159. URL https://aclanthology.org/P16-1159.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7909–7913. IEEE, 2020.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4252–4261, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.328. URL https://aclanthology.org/2021.acl-long.328.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 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)*, pp. 6209–6213, 2021b. doi: 10.1109/ICASSP39728.2021.9415058.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 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)*, pp. 1488–1499, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.105. URL https://aclanthology.org/2022.acl-long.105.

Jörgen Valk and Tanel Alumäe. VOXLINGUA107: A dataset for spoken language recognition. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pp. 652–658. IEEE, 2021. doi: 10.1109/SLT48900.2021.9383459. URL https://doi.org/10.1109/SLT48900.2021.9383459.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020a.

Changhan Wang, Anne Wu, and Juan Miguel Pino. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310, 2020b. URL https://arxiv.org/abs/2007.10310.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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)*, pp. 993–1003, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. Large-scale self- and semi-supervised learning for speech translation. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (eds.), *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pp. 2242–2246. ISCA, 2021b. doi: 10.21437/Interspeech.2021-1912. URL https://doi.org/10.21437/Interspeech.2021-1912.

Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3544–3552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.326. URL https://aclanthology.org/2020.acl-main.326.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proc. of AAAI*, volume 34, pp. 9161–9168, 2020c.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3728–3738, Online, July 2020d. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.344. URL https://aclanthology.org/2020.acl-main.344.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. In *Proc. of INTERSPEECH*, pp. 2625–2629, 2017.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3612–3621, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1397. URL https://aclanthology.org/D18-1397.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2619–2630, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.204. URL https://aclanthology.org/2021.acl-long.204.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. Bilingual mutual information based adaptive training for neural machine 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)*, pp. 511–516, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.65. URL https://aclanthology.org/2021.acl-short.65.

Rong Ye, Mingxuan Wang, and Lei Li. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*, August 2021. URL https://www.isca-speech.org/archive/pdfs/interspeech_2021/ye21_interspeech.pdf.

Rong Ye, Mingxuan Wang, and Lei Li. 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*, pp. 5099–5113, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.376.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148.

Biao Zhang, Barry Haddow, and Rico Sennrich. Revisiting end-to-end speech-to-text translation from scratch. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26193–26205. PMLR, 2022a. URL https://proceedings.mlr.press/v162/zhang22i.html.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. Conditional bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2377–2389, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.169. URL https://aclanthology.org/2022.acl-long.169.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4334–4343, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1426. URL https://aclanthology.org/P19-1426.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proc. of ICML*, 2021.

A Statistics of all datasets

Table 3: Statistics of all datasets. #sents refers to the number of parallel sentence pairs.

	ST (MuST-C)		External MT	
En→	hours	#sents	name	#sents
De	408	234K	WMT16	3.9M
Fr	492	280K	WMT14	31.2M
Es	504	270K	WMT13	14.2M
Ro	432	240K	WMT16	0.6M
Ru	489	270K	WMT16	1.9M
It	465	258K	OPUS100	0.7M
Pt	385	211K	OPUS100	0.7M
Nl	442	253K	OPUS100	0.7M

B The Choice of Decay Parameter in Scheduled Sampling

In scheduled sampling, the probability of selecting the ground truth word p^* keeps decreasing during training as the function in Equation (10). Here, the hyper-parameter μ is used to control the shape of the function. As μ increases, the probability p^* decreases more slowly, and vice versa. We investigate the impact of μ in Figure 8, and find that (1) the model performs worse when p^* drops too quickly, and (2) when μ is within a reasonable range, there is not much impact on the final BLEU score. We use $\mu = 15$ in our experiments.

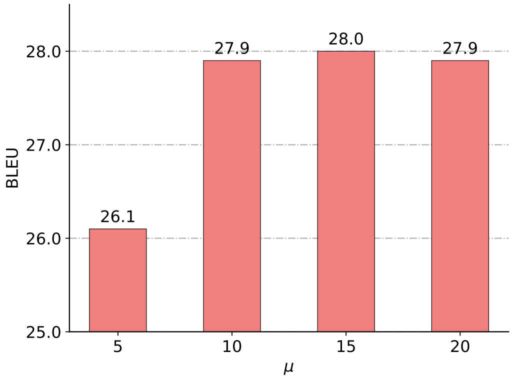

Figure 8: BLEU scores on MuST-C En→De `dev` set (expanded setting) with different μ. Here token-level adaptive training is not used for training.

C Discussion about the Training Speed

During training, our approach requires an additional forward pass to select predicted words compared with baseline, which will impair the speed of training. Practically, we find the training time for 1 epoch of **Cress** is 1.12 times longer than **MTL**, which is actually negligible. This is because the step of selecting predicted words is fully parallel and has no gradient calculation, which does not incur a significant time overhead.

D IMPACT OF DIFFERENT ACOUSTIC ENCODERS

Our model is composed of an acoustic encoder and a translation model. To investigate the impact of different acoustic encoders, we also conduct experiments using Wav2vec 2.0[8] (Baevski et al., 2020) as the acoustic encoder. As shown in Table 4, we find that (1) HuBERT performs slightly better than Wav2vec 2.0 with an improvement of 0.5 BLEU, and (2) our proposed **CRESS** achieves consistent improvements with different acoustic encoders.

Table 4: BLEU scores on MuST-C En→De `tst-COMMON` set (expanded setting) with different acoustic encoders.

Acoustic Encoder	**MTL**	**CRESS**
HuBERT (Hsu et al., 2021)	27.5	**29.4**
Wav2vec 2.0 (Baevski et al., 2020)	27.0	**28.9**

E CHRF++ SCORES ON MUST-C DATASET

We also report ChrF++ score (Popović, 2017) using sacreBLEU toolkit[9] on MuST-C dataset in Table 5. We observe that **CRESS** outperforms **MTL** with 1.4 ChrF++ improvement in the base setting and 1.0 ChrF++ improvement in the expanded setting.

Table 5: ChrF++ scores on MuST-C `tst-COMMON` set. The external MT datasets are only used in the expanded setting. * and ** mean the improvements over **MTL** baseline is statistically significant ($p < 0.05$ and $p < 0.01$, respectively).

Models	**ChrF++**								
	En→De	En→Fr	En→Es	En→Ro	En→Ru	En→It	En→Pt	En→Nl	Avg.
Base setting (w/o external MT data)									
MTL	52.4	60.4	56.4	50.9	41.7	52.6	57.3	56.1	53.5
CRESS	**54.0****	**62.0****	**57.6****	**52.4****	**43.1****	**53.8****	**58.5****	**57.6****	**54.9**
Expanded setting (w/ external MT data)									
MTL	54.9	62.6	58.6	51.9	44.2	53.4	57.9	56.9	55.0
CRESS	**56.1****	**63.7****	**58.9***	**53.1****	**44.5***	**54.2****	**59.3****	**58.3****	**56.0**

F RESULTS ON COVOST 2 EN→DE

We also conduct experiments on CoVoST 2 (Wang et al., 2020b) to examine the performance of our approach on large datasets. CoVoST 2 is a large-scale multilingual speech translation corpus which covers translations from 21 languages into English and from English into 15 languages. It is one of the largest open ST dataset available currently. In this paper, we evaluate our approach on the En→De direction, which contains 430 hours of speech with annotated transcriptions and translations. We use `dev` set for validation and `test` set for evaluation.

We use the same pre-processing, model configuration, and hyper-parameters as MuST-C (see details in Section 5.2). The results are shown in Table 6. First, we find our **CRESS** significantly outperforms the **MTL** baseline, with 1.8 BLEU improvement in the base setting and 1.6 BLEU improvement in the expanded setting, which demonstrates the effectiveness and generalization capability of our method across different datasets, especially on the large-scale dataset. Second, our result is competitive with previous methods, although they use larger audio datasets ($\geq$60K hours) and larger model size ($\geq$300M), while we only use 960 hours of audio data and 155M model parameters.

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

[9] sacreBLEU signature: nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0

Table 6: BLEU scores on CoVoST 2 En→De `test` set. LS-960: LibriSpeech (Panayotov et al., 2015) (960 hours). LV-60K: Libri-Light (Kahn et al., 2020) (60K hours). VP-400K: VoxPopuli (Wang et al., 2021a) (372K hours). MLS: Multilingual LibriSpeech (Pratap et al., 2020) (50K hours). CV: CommonVoice (Ardila et al., 2020) (7K hours). VL: VoxLingua107 (Valk & Alumäe, 2021) (6.6K hours). BBL: BABEL (Gales et al., 2014) (1K hours).

Models	**Audio Datasets**	**#Params**	**BLEU**
wav2vec-2.0 (LS-960) (Wang et al., 2021b)	LS-960	300M	20.5
wav2vec-2.0 (LV-60K) (Wang et al., 2021b)	LV-60K	300M	25.5
wav2vec-2.0 + Self-training (LV-60K) (Wang et al., 2021b)	LV-60K	300M	**27.1**
LNA (Joint Training) (Li et al., 2021)	LV-60K	1.05B	25.8
SLAM-TLM (Bapna et al., 2021)	LV-60K	600M	**27.5**
XLS-R (0.3B) (Babu et al., 2022)	VP-400K, MLS, CV, VL, BBL	317M	23.6
XLS-R (1B) (Babu et al., 2022)	VP-400K, MLS, CV, VL, BBL	965M	26.2
XLS-R (2B) (Babu et al., 2022)	VP-400K, MLS, CV, VL, BBL	2162M	**28.3**
MTL (base setting)	LS-960	**155M**	21.4
CRESS (base setting)	LS-960	**155M**	23.2 (+1.8)
MTL (expanded setting)	LS-960	**155M**	25.1
CRESS (expanded setting)	LS-960	**155M**	**26.7** (+1.6)

G Performance of Text Translation

Since our method is built upon the multi-task learning framework, we also report the performance of MT task. As shown in Table 7, our method not only brings improvements to ST, but also gives a slight average boost of 0.3 BLEU to MT. We suggest that this may be due to the effect of regularization. More importantly, we find that the performance gap between ST and MT for **CRESS** is significantly reduced compared to the **MTL** baseline (6.0→5.0), which further demonstrates that the improvement in ST performance is mainly due to the effective reduction of the modality gap.

Table 7: BLEU scores of both ST and MT on MuST-C `tst-COMMON` set (expanded setting). Δ indicates the average gap in BLEU between ST and MT.

Models	**Task**	**BLEU**									
		En→De	En→Fr	En→Es	En→Ro	En→Ru	En→It	En→Pt	En→Nl	Avg.↑	Δ ↓
MTL	ST	27.7	38.5	32.8	24.9	19.0	26.5	32.0	30.8	29.0	6.0
	MT	33.5	**46.6**	**38.3**	30.9	22.1	33.0	38.6	36.7	35.0	
CRESS	ST	**29.4**	**40.1**	**33.2**	**26.4**	**19.7**	**27.6**	**33.6**	**32.3**	**30.3**	**5.0**
	MT	**34.1**	**46.6**	38.1	**31.1**	**22.4**	**33.3**	**39.5**	**37.6**	**35.3**	