Efficient CTC Regularization via Coarse Labels for End-to-End Speech Translation

Biao Zhang[1]   **Barry Haddow**[1]   **Rico Sennrich**[2,1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computational Linguistics, University of Zurich
B.Zhang@ed.ac.uk, bhaddow@ed.ac.uk, sennrich@cl.uzh.ch

Abstract

For end-to-end speech translation, regularizing the encoder with the Connectionist Temporal Classification (CTC) objective using the source transcript or target translation as labels can greatly improve quality metrics. However, CTC demands an extra prediction layer over the vocabulary space, bringing in non-negligible model parameters and computational overheads, although this layer is typically not used for inference. In this paper, we re-examine the need for genuine vocabulary labels for CTC for regularization and explore strategies to reduce the CTC label space, targeting improved efficiency without quality degradation. We propose coarse labeling for CTC (CoLaCTC), which merges vocabulary labels via simple heuristic rules, such as using truncation, division or modulo (MOD) operations. Despite its simplicity, our experiments on 4 source and 8 target languages show that CoLaCTC with MOD particularly can compress the label space aggressively to 256 and even further, gaining training efficiency ($1.18\times \sim 1.77\times$ speedup depending on the original vocabulary size) yet still delivering comparable or better performance than the CTC baseline. We also show that CoLaCTC successfully generalizes to CTC regularization regardless of using transcript or translation for labeling.

1 Introduction

Developing techniques to support the translation from a source-language audio to a target-language text directly, or end-to-end (E2E) speech translation (ST), has attracted increasing attention recently due to its potential of reducing translation latency and avoiding error propagation (Duong et al., 2016; Bérard et al., 2016). However, solving this task is non-trivial because of the speech-text modality gap: one word corresponds to a stochastic sequence of speech signals that vary greatly across speakers and over contexts, which increases the learning difficulty. Recent progress on E2E ST mainly focuses on bridging this gap through the encoder-decoder framework from diverse perspectives (Di Gangi et al., 2019; Salesky et al., 2019; Zhang et al., 2020; Wang et al., 2020b; Han et al., 2021; Zheng et al., 2021).

CTC regularization is such an approach that facilitates the modeling of translation by aligning speech representations from the encoder with discrete labels dynamically via the lens of the Connectionist Temporal Classification (CTC) objective (Graves et al., 2006b). Bahar et al. (2019) first examined the use of the source transcript as discrete labels, improving translation quality consistently across various ST settings; Zhang et al. (2022) further discovered that using the target translation as labels instead can also be surprisingly effective although speech-translation pairs arguably violates CTC's monotonicity prerequisite. Nevertheless, these successes come at the cost of increased computational overheads and model parameters because CTC demands an extra prediction layer over its label space for probability estimation and this space is often huge – traditionally the source or target vocabulary size (Gaido et al., 2020). We thus explore strategies to achieve the best of both worlds, i.e., improving the efficiency of CTC regularization without hurting its performance.

We address this problem by reexamining the need for genuine vocabulary labels for CTC. In contrast to CTC-based generation (Graves et al., 2006b), the prediction layer in CTC regularization of ST is discarded at inference. In other words, sticking to genuine labels is computationally unnecessary. Since the large label space of CTC is a crucial bottleneck hindering training efficiency, we explore ways of reducing it. We propose **Co**arse **La**beling for CTC (CoLaCTC) that manipulates this space by merging vocabulary labels based on simple heuristic rules. Concretely, we map the source or target vocabulary to a pseudo label space subject to some predefined size using simple op-

Method		Mapping	# Labels	ID Sequence
Genuine Labels		`f(z) = z`	V	`0,1,2,3,4,5,6,7,8`
CoLaCTC	Truncation	`f(z) = min(z, L-1)`	L	`0,1,2,2,2,2,2,2,2`
	Modulo	`f(z) = z % L`	L	`0,1,2,0,1,2,0,1,2`
	Division	`f(z) = ⌊z / (V/L)⌋`	L	`0,0,0,1,1,1,2,2,2`
	Log-Scaling	`f(z) = ⌊log(max(z,1)) * L/log(V)⌋`	L	`0,0,0,1,1,2,2,2,2`

Table 1: Overview of different mappings for CoLaCTC. V and L denote the original vocabulary size and the specified coarse label size, respectively, and $V \gg L$. z is the vocabulary-space token ID starting from 0. The ID sequence is just for a toy example, and we set $V = 9, L = 3$ for illustration. $f(\cdot)$ shows the mapping function. CoLaCTC adopts different operations to reduce the label space from V to L.

erations, such as *truncation*, *modulo*, *division* and *log-scaling* as shown in Table 1.

Despite the label space being transformed, the generated coarse labels still maintain a strong correlation with their vocabulary counterparts, ensuring their informativeness for representation learning. We rigorously examined our method on the MuST-C (Di Gangi et al., 2019) and the Multilingual TEDx (Salesky et al., 2021) benchmarks, covering 4 source languages and 8 target languages. Across diverse settings, CoLaCTC successfully achieves comparable or even better translation performance than the CTC baseline but with significantly improved training efficiency (up to $1.77\times$ speedup depending on the original vocabulary size). Our main contributions are summarized below:[1]

- We propose coarse labeling for CTC regularization which offers a mechanism to decouple the CTC label size from the vocabulary size; with CoLaCTC, a CTC-regularized model can be trained nearly as fast as a non-CTC model.

- We compare two types of CTC regularization for ST, i.e., using transcript or translation for labeling, and show that transcript performs better *when it is available*.

- CoLaCTC delivers promising performance on 4 source and 8 target languages, and also generalizes to both types of CTC regularization.

- Our empirical analysis reveals that CoLaCTC benefits translation similarly to the CTC baseline on different aspects, including homophone translation, and seems to improve the contextualization of speech representations.

2 Related Work

Solving E2E ST requires techniques to mitigate the speech-text modality gap. One way is to develop advanced architectures integrating speech-specific characteristics to the encoder, such as locality modeling for the self-attention (Di Gangi et al., 2019; Gulati et al., 2020) and adaptive speech representation grouping (Salesky et al., 2019; Liu et al., 2020; Zhang et al., 2020). Another way is to leverage knowledge from other languages and/or tasks, including automatic speech recognition (ASR) and machine translation (MT) based multi-task modeling (Anastasopoulos and Chiang, 2018; Dong et al., 2021; Du et al., 2021), cross-lingual transfer learning (Inaguma et al., 2019; Di Gangi et al., 2019; Li et al., 2021), and large-scale weakly, semi- and self-supervised pretraining (Schneider et al., 2019; Ao et al., 2021; Bapna et al., 2022). Our method contributes to E2E ST by accelerating CTC regularization with coarse labels, and is theoretically orthogonal to all the techniques aforementioned. In this study, we mainly focus on bilingual ST using triplet data alone, and leave the exploration of how our method is compatible with other setups to future work.

CTC was first proposed to handle the sequence mismatch problem between acoustic features and transcript tokens, and has been widely applied to ASR (Graves et al., 2006b,a; Chan et al., 2016) and other tasks where the input sequence is longer than the output and their alignment is monotonic (Niu and Mak, 2020; Cai et al., 2022). Recent studies also show promising results when applying CTC to non-monotonic tasks, specifically to non-autoregressive ST and MT (Libovický and Helcl, 2018; Saharia et al., 2020; Gu and Kong, 2021; Chuang et al., 2021). All these methods treat the prediction layer in CTC as a generator, used to predict final outputs. By contrast, Kim et al. (2017) used CTC as an auxiliary objective to improve ASR. In E2E ST, Liu et al. (2020), Gaido et al. (2021), Xu et al. (2021) and Dong et al. (2021) leveraged CTC to compress speech representations to bridge

---

[1]Source code: `https://github.com/bzhangGo/zero`.

the modality gap; Bahar et al. (2019) and Gaido et al. (2022) explored CTC as a regularizer using source transcript as labels, showing encouraging performance although the prediction layer is not used at inference; Zhang et al. (2022) further investigated E2E ST without transcript, and found that CTC regularization with translation as labels also works. Our study follows CTC regularization and extends it with CoLaCTC to address its training inefficiency issue. As far as we know, exploring coarse labels for CTC has never been investigated before, at least on ST.

3 Background: CTC Regularization

CTC regularization improves the encoder-decoder based E2E ST by adding a CTC regularizer to the conventional translation loss (Bahar et al., 2019; Zhang et al., 2022). Formally, given a *(source speech, target translation)* pair denoted as (X, Y) respectively, E2E ST with CTC regularization is optimized via the following interpolated objective:[2]

$$\underbrace{(1 - \lambda)\mathcal{L}^{\text{MLE}}(Y|\mathbf{Y})}_{\text{translation loss}} + \underbrace{\lambda\mathcal{L}^{\text{CTC}}(Z|\mathbf{X})}_{\text{regularization}}, \quad (1)$$

where λ is a hyperparameter balancing different sub-objectives. $\mathbf{X} \in \mathbb{R}^{|X| \times d}$ and $\mathbf{Y} \in \mathbb{R}^{|Y| \times d}$ denote the encoder output (or speech representation) and the decoder output, respectively. Z is the label sequence for CTC, which is often either the source transcript or the target translation. $|\cdot|$ indicates the sequence length and d is the model dimension.

The translation loss $\mathcal{L}^{\text{MLE}}$ aims at maximizing the likelihood of observed training instances. Often, we decompose the likelihood token-wise in accordance with the autoregressive generation

$$\mathcal{L}^{\text{MLE}}(Y|\mathbf{Y}) = -\sum_t \log p(y_t|y_{<t}), \quad (2)$$

$$\text{with} \quad p(y_t|y_{<t}) \sim \text{softmax}\left(\mathbf{W}^{\text{MLE}}\mathbf{y}_t\right), \quad (3)$$

where y_t stands for the t-th target token. $\mathbf{y}_t \in \mathbb{R}^d$ is the t-th row of $\mathbf{Y}$, representing the translation prefix $y_{<t} = \{y_1, \ldots, y_{t-1}\}$. $\mathbf{W}^{\text{MLE}} \in \mathbb{R}^{V^{\text{MLE}} \times d}$, also called softmax embedding, is a trainable parameter, and V^{MLE} is the target vocabulary size. E2E ST uses this embedding to estimate the translation probability of each target word as shown in Eq. (3)[3]. At inference, the emitted probability offers direct evidence to search translation candidates.

By contrast, the regularization term $\mathcal{L}^{\text{CTC}}$ encourages the dynamic alignment of speech representations ($\mathbf{X}$) with their corresponding discrete labels (Z) through the CTC algorithm. CTC regularization also maximizes the likelihood. But different from the token-by-token formulation in Eq. (2), CTC estimates the likelihood by marginalizing over all valid mappings between the input and output sequence (Graves et al., 2006b)

$$\mathcal{L}^{\text{CTC}}(Z|\mathbf{X}) = -\log \sum_{A \in \Gamma(Z)} \prod_k p(a_k|\mathbf{x}_k), \quad (4)$$

$$\text{with} \quad p(a_k|\mathbf{x}_k) \sim \text{softmax}\left(\mathbf{W}^{\text{CTC}}\mathbf{x}_k\right), \quad (5)$$

where $\Gamma(Z)$ denotes the set of all valid aligned sequences. The probability of each aligned label a_k in the sequence A is estimated by a prediction layer based on the corresponding speech representation $\mathbf{x}_k$ as shown in Eq. (5). $\mathbf{W}^{\text{CTC}} \in \mathbb{R}^{V^{\text{CTC}} \times d}$ is the prediction parameter, and $V^{\text{CTC}} = V + 1$ denotes the CTC label size. When the source transcript (target translation) is used as labels, V is the source (target) vocabulary size; the extra label is for the special *blank* symbol. We refer readers to Graves et al. (2006b) for more details. Note that this predication layer will be discarded after training when CTC is purely used for regularization.

Previous studies have examined using either source transcript (Bahar et al., 2019; Gaido et al., 2022) or target translation (Zhang et al., 2022) as labels for CTC regularization, but separately. In Section 5.1, we will compare these two types of CTC regularization under the same setup.

4 Coarse Labeling for CTC

Unfortunately, CTC regularization suffers from training inefficiency. Despite CTC being efficiently addressed via dynamic programming, its prediction layer in Eq. (5) is unavoidable. This layer introduces considerable computational overhead due to scaling linearly with the vocabulary size and also brings in large number of model parameters particularly when $V \neq V^{\text{MLE}}$. How to improve the efficiency, and save model parameters while retaining the performance is the focus of our study.

We draw inspiration from the fact that the prediction layer in CTC regularization is not used for inference. Thus, sticking to genuine vocabulary labels is unnecessary technically. If we could design pseudo labels in a reduced label space as alternatives to the genuine ones, that would address the

---

[2]Note we use X, Y to denote the input, and their bold variants to denote the learned hidden representations.

[3]We drop the bias term for clarity.

inefficiency issue. Following this intuition, we propose coarse labeling for CTC (CoLaCTC) which formulates the pseudo label generation process as a vocabulary mapping:

$$f(z) : \mathbb{N}_{[0...V-1]} \mapsto \mathbb{N}_{[0...L-1]}, \qquad (6)$$

where z denotes the original vocabulary ID. L is a hyperparameter specifying the label size, and we often set $L \ll V$. This transformation decouples the label size of CTC from the vocabulary size, offering flexibility to optimize the training efficiency. One assumption behind such formulation is that the success of CTC regularization mainly comes from the inductive biases of the CTC algorithm rather than the genuineness of its labels[4], which we verified empirically through experiments.

Eq. (6) merges a set of vocabulary labels into one label according to $f(\cdot)$. Potential mappings are many, such as grouping semantically similar words or considering phonetic similarity. But these linguistically inspired approaches often lack freedom in manipulating the label space (L). Instead, we adopt the following heuristic methods (as shown in Table 1):[5]

Truncation (TRU) Learning speech representations for infrequent items is often difficult, so we merge all vocabulary labels except for the top-$(L-1)$ in TRU

$$f(z) = \min(z, L-1), \qquad (7)$$

with the hope that those frequent items could provide informative clues for CTC.

Modulo (MOD) Nevertheless, infrequent items might carry crucial content information. Instead of naively collapsing them, MOD merges diverse labels of varying frequencies based on a fixed interval

$$f(z) = z \pmod{L}. \qquad (8)$$

Division (DIV) Different from items with varying frequencies, items of similar frequency often share similar linguistic properties. We explore

---

[4]The inductive biases include the modeling of local structures for speech, word boundary identification and label-guided speech representation learning, etc.

[5]Note that we followed the standard practice and ranked the items in our vocabulary based on their frequency. In this study, the items are (sub)words.

this in DIV which merges labels of similar ranks uniformly

$$f(z) = \lfloor z * {}^L/_V \rfloor, \qquad (9)$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

Log-Scaling (LOG) One drawback of DIV is that the distribution of its coarse labels becomes badly skewed. To offset this problem, we further study a non-linear, log-scaled transformation, LOG

$$f(z) = \lfloor \log(\max(z, 1)) * {}^L/_{\log(V)} \rfloor. \quad (10)$$

Note we intentionally use simple operations to keep the simplicity of CoLaCTC. All the above operations are trivial to implement.

Although labels generated by these operations become linguistically less meaningful, they still keep a strong correlation with their genuine vocabulary counterparts. We expect this correlation could ensure the informativeness of each coarse label and further facilitate the generalization of CoLaCTC to CTC regularization. We compare different operations via experiments.

5 Experiments

Setup We work on two benchmarks, Multilingual TEDx (Salesky et al., 2021) and MuST-C (Di Gangi et al., 2019), covering 4 source and 8 target languages. MuST-C (v1) is an English-audio based multilingual corpus, including translations from English (En) to 8 languages: German (De), Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Romanian (Ro) and Russian (Ru). The training data for each language pair has ∼452 hours with about 252K utterances on average, and we use the given dev and tst-COMMON splits as the dev and test set, respectively. In contrast, Multilingual TEDx is a multi-source and multi-target ST corpus, containing audios in diverse languages, although its scale is relatively small. We regard this benchmark as a testbed to examine the applicability of our method to audios other than English. We report results on 6 translation directions, i.e., Es-En, Es-Pt, Fr-En, Pt-En, Fr-Es and Fr-Pt. The training data of different language pairs ranges from 25 hours (16K utterances, Fr-Pt) to 69 hours (39K utterances, Es-En), and we use the official dev and test sets for experiments.
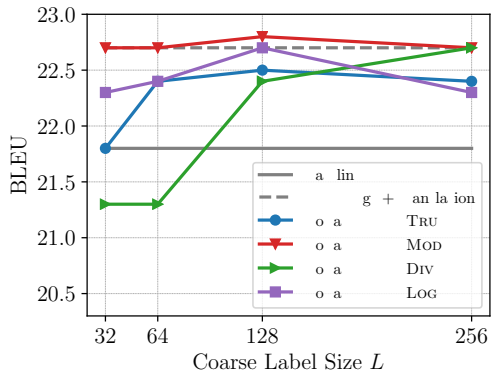
Figure 1: Translation results of different CoLaCTC methods on the MuST-C En-De test set as a function of the coarse label size L. *Reg.*: short for regularization. Here we use **translation** labels for CTC regularization.
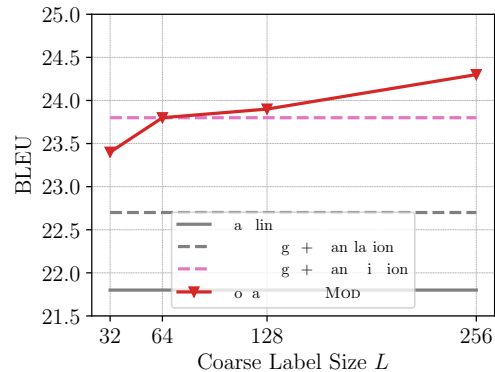

Figure 2: Translation results for CoLaCTC with MOD on the MuST-C En-De test set as a function of L. Here we use **transcription** labels for CTC regularization.

System	#Param	BLEU	Speedup
Baseline	46.1M	21.8	1.39×
CTC Reg. + translation	47.9M	22.7	1.00×
+ CoLaCTC	46.2M	22.7	1.39×
+ share parameters	46.1M	22.4	0.97×
CTC Reg. + transcription	47.5M	23.8	1.00×
+ CoLaCTC	46.2M	24.3	1.31×

Table 2: Test results of different systems on MuST-C En-De. $L = 256$ for CoLaCTC. *#Param*: the number of parameters. *share parameters*: share parameters between $\mathbf{W}^{\text{MLE}}$ and $\mathbf{W}^{\text{CTC}}$. We perform three runs (50 steps each) to evaluate the training speedups on GeForce GTX TITAN X.

We focus on bilingual ST and adopt the E2E ST model following Zhang et al. (2022) which concatenates neighboring frames for downsampling followed by a variant of Transformer for translation. We set $\lambda = 0.3$ for CTC regularization. We evaluate the translation quality using (Sacre)BLEU (Post, 2018).[6] We didn't perform any filtering to the test set. All models are implemented in *Tensorflow*, and trained from scratch without any ASR or MT pretraining. We refer readers to Appendix A for details on data preprocessing and training.

5.1 Analysis on MuST-C En-De

Coarse label size matters, and MOD performs the best. CoLaCTC depends on not only the mapping function selected, but also the coarse label size specified. In general, CoLaCTC with a larger label size produces coarse labels closer to the genuine ones, thus behaving more robustly. We first perform ablations for CoLaCTC with the target translation as labels, where only speech-translation pairs are used at training (Zhang et al., 2022).

We vary L from 32 to 256, and show the results in Figure 1. Different mapping functions have different properties and also show different behaviors, where the label size yields profound impacts. When L is small, DIV performs the worst, followed by LOG and TRU while MOD performs the best. With the increase of L, the performance difference between different mappings narrows. Label size matters, but the optimal size varies for different mappings. Under different settings, MOD performs the best and is most robust, nearly dominating the others. We next mainly study MOD for CoLaCTC.

CoLaCTC performs comparably to the CTC baseline. Figure 1 also shows that CoLaCTC delivers comparable results to the CTC baseline when proper L is applied; with MOD even across all tested L. Note both methods significantly outperform the vanilla baseline without CTC regularization. This demonstrates that the genuineness of CTC labels matters less for CTC regularization and that our strategy – generating coarse labels in a reduced space – is feasible.

Transcript is more effective than translation as labels for CTC regularization. Despite being effective, using translation as labels for CTC regularization violates the monotonic assumption required by CTC. CTC with transcripts is more established (Bahar et al., 2019; Gaido et al., 2022). We thus compare these two types of CTC regularization and explore how CoLaCTC generalizes.

Figure 2 shows that using transcript instead yields substantial quality improvements (+1.0 BLEU), and that CoLaCTC with MOD generalizes to both settings successfully. Still, the genuineness of CTC labels matters less than their origin does!
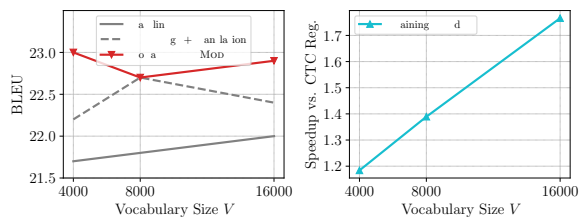
---

[6]Signature: *BLEU+c.md+#ref.1+s.exp+tok.13a+v.1.4.14*

Figure 3: Test results for different methods on MuST-C En-De when varying the target vocabulary size V. We adopt **translation** labels for CTC regularization. $L = 256$ for CoLaCTC.
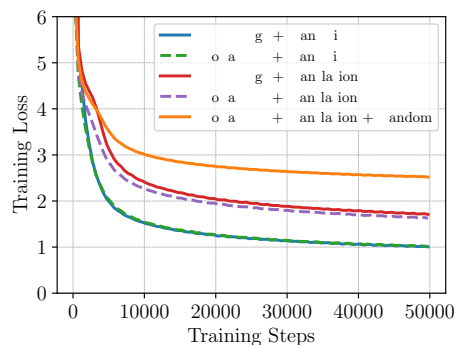

Figure 4: Training loss as a function of training steps for different methods on MuST-C En-De. $L = 256$ for Co-LaCTC. *Random* indicates using fully random labels (i.e. `f(z) = randint(0, L)`) with a label size of 256.

We observe that using 256 coarse labels works well for CTC regularization under different settings. We set $L = 256$ for the following experiments.

CoLaCTC saves model parameters and greatly improves training efficiency. CTC regularization suffers from inefficiency, which increases model parameters by about 4% and slows the training by 39% as shown in Table 2. We try to solve this problem by sharing parameters between the CTC prediction layer and the softmax output layer when $V = V^{\text{MLE}}$ is the target vocabulary size. Unfortunately, this hurts quality and helps the training efficiency little (+ *share parameters*). By contrast, CoLaCTC nearly recovers the efficiency sacrificed by CTC regularization, running as fast as the vanilla baseline but still retaining quality improvements. Besides, CoLaCTC performs similarly well with different label sequences.

We also note that the degree of inefficiency depends on the computational framework used. We re-tested different methods with PyTorch, where CTC regularization causes a 10% decrease in training speed, much smaller than 39%. However, the conclusion that CTC regularization leads to more trainable parameters and slower running speed, and that CoLaCTC overcomes this issue, still holds.

CoLaCTC performs robustly over different vocabulary sizes; larger V yields higher speedups. Apart from the coarse label size, the target vocabulary size V also affects CTC regularization. Larger vocabulary shortens the target sequence but increases the CTC label space. Figure 3 shows the impact of V on CoLaCTC. Translation performance is highly sensitive to the vocabulary size. Using CTC regularization delivers consistent quality gains against the vanilla baseline, and CoLaCTC shows promising robustness, matching and even outperforming CTC regularization with genuine labels. Regarding training efficiency, the speedup of CoLaCTC should scale linearly with V in the-

ory when L is fixed. Figure 3 confirms this where CoLaCTC achieves higher speedups with larger vocabulary sizes. Particularly, the speedup reaches **1.77×** when V is 16K, a substantial improvement.

CoLaCTC doesn't hurt the trainability of ST models. Would CoLaCTC increase the learning difficulty, which likely reduces performance? Figure 4 shows that 1) CoLaCTC shows similar convergence to the CTC baseline using either transcript or translation for labelling; 2) the model using transcript as labels converges faster and to a better local optima than the counterpart using translation, which also explains the results in Figure 2; 3) random coarse labels result in inferior convergence due to their unpredictable nature.

The vocabulary order of genuine labels has limited impact on CoLaCTC. As shown in Eq. (6), the coarse labeling in CoLaCTC highly relies on the order of original labels in the vocabulary. This ordering encodes word frequency, which might offer crucial clues to CoLaCTC and explain its success. We examine this by randomly shuffling the vocabulary, thus the original vocabulary ID is randomly changed and the ordering information is eliminated. With this shuffled vocabulary, Co-LaCTC (MOD) achieves a BLEU score of 23.9 on MuST-C En-De test set, matching the performance of the CTC baseline (23.8) although underperforming the original CoLaCTC (24.3). CoLaCTC benefits from the order information but still achieves promising performance without it.

5.2 Results on Other Languages

CoLaCTC achieves great performance for translation out of English. Table 3 summarizes the results of CoLaCTC on other MuST-C translation

System	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg
ESPnet-ST (Inaguma et al., 2020)†	22.9	28.0	32.8	23.8	27.4	28.0	21.9	15.8	25.1
Contextual Modeling (Zhang et al., 2021)	22.9	27.3	32.5	23.1	26.0	27.1	23.6	15.8	24.8
Fairseq-ST (Wang et al., 2020a)†	22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3	24.8
NeurST (Zhao et al., 2021)	22.8	27.4	33.3	22.9	27.2	28.7	22.2	15.1	24.9
Wav2Vec-Transformer (Han et al., 2021)	22.3	28.7	34.3	24.2	28.2	29.3	22.4	15.8	25.7
E2E-ST-JT (Du et al., 2021)†	23.1	27.5	32.8	23.6	27.8	28.7	22.1	14.9	25.1
E2E-ST-TDA (Du et al., 2021)†	24.3	28.3	34.6	24.2	28.7	30.3	23.4	15.9	26.2
Baseline	21.8	27.3	32.3	22.5	26.6	27.5	21.8	14.7	24.3
CTC regularization + target translation	22.7	28.1	33.4	23.2	26.9	28.3	22.6	15.4	25.1
+ CoLaCTC	22.7	27.9	33.3	23.7	27.1	28.0	22.4	15.9	25.1
CTC regularization + source transcription	23.8	28.6	33.9	24.3	28.3	29.3	23.3	16.3	26.0
+ CoLaCTC	24.3	28.4	34.5	24.6	28.1	28.8	23.3	16.6	**26.1**

Table 3: BLEU of different systems on MuST-C tst-COMMON. *Avg*: average score over different language pairs (translation is always out-of English). †: systems that might perform filtering to the test set, meaning results are not necessarily comparable. *Baseline*: the model without CTC regularization; $L = 256$ and MOD for CoLaCTC.

System	Es-En	Es-Pt	Fr-En	Pt-En	Fr-Es	Fr-Pt	Avg
Bilingual Cascades (Salesky et al., 2021)	15.5	23.3	17.2	16.1	17.8	12.2	17.0
Bilingual E2E ST (Salesky et al., 2021)	7.0	12.2	8.9	8.1	10.6	7.9	9.1
Multilingual E2E ST (Salesky et al., 2021)	12.3	17.4	12.0	12.0	13.6	13.2	13.4
Baseline	11.6	13.3	7.6	8.5	6.1	1.9	8.2
CTC regularization + target translation	13.0	18.2	12.2	11.4	11.5	6.1	12.1
+ CoLaCTC	13.3	19.0	12.1	12.0	11.2	5.0	12.1
CTC regularization + source transcription	18.0	23.0	19.3	17.8	19.8	13.8	**18.6**
+ CoLaCTC	17.8	23.1	19.9	17.5	19.6	13.3	18.5

Table 4: BLEU Scores on Multilingual TEDx test sets. $L = 256$ and MOD for CoLaCTC.

directions. The performance of CoLaCTC varies across different languages with both positive and negative gains. But overall, CoLaCTC is on par with its CTC baselines and largely outperforms the vanilla baseline without CTC regularization. On average, CoLaCTC delivers a BLEU score of 25.1 and 26.1 when used with translation and transcription labels, respectively, which also surpasses many strong previous studies (Inaguma et al., 2020; Zhao et al., 2021; Zhang et al., 2021).

Joint training with CTC regularization is preferable to the traditional pretraining-finetuning paradigm for E2E ST. The current *de facto* standard for training an E2E ST model is to firstly initialize it with a pretrained ASR encoder and/or MT decoder and then finetune it on ST data. Despite its effectiveness, this pipeline paradigm often consumes longer training time and inevitably complicates the optimization procedure. In contrast, joint training with CTC regularization is technically simpler and delivers comparable and even better results as shown in Table 3, echoing with Gaido et al. (2022). Note that we also re-implemented the pipeline baseline using our in-house codebase, which achieves 22.9 BLEU on MuST-C En-De, far

below the joint training with transcript (23.8).

Since CoLaCTC solves the inefficiency issue for CTC regularization, we would recommend using the joint training as the new standard for E2E ST, especially when only triplet training data is used.

CoLaCTC generalizes to ST settings other than English audios. The above results for CoLaCTC are multilingual, but all use English audio from TED on the source side. To demonstrate generalization across different source languages, we conduct experiments on Multilingual TEDx and work on ST for Es, Fr and Pt. Table 4 shows that CoLaCTC generalizes well to other source languages. In addition, CTC regularization performs much better on this benchmark, substantially outperforming the vanilla baseline by 10.4 BLEU, matching the performance of multilingual ST reported by Salesky et al. (2021). We ascribe this success to the small scale of Multilingual TEDx where regularization techniques, like CTC Reg., often work better.

6 Discussion

The promising performance of CoLaCTC inspires us to further explore why coarse labels could work for CTC regularization. Analyzing the underlying
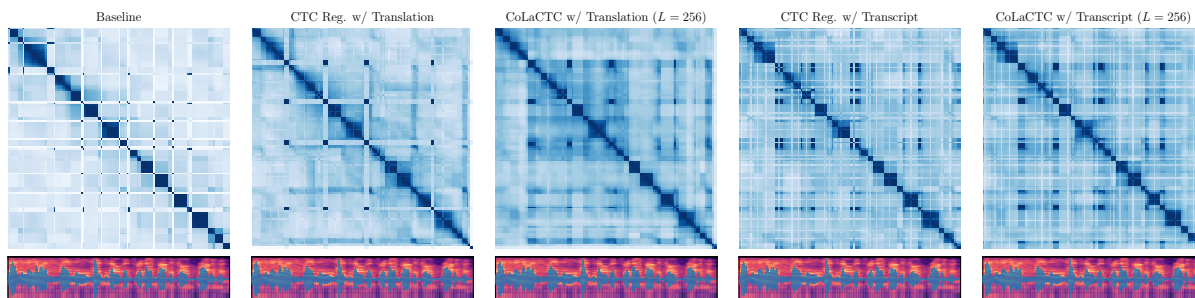
Figure 5: Visualization of the cosine similarity of the final-layer speech encoding for different methods. *Top*: cosine similarity where darker color shows higher similarity; *Bottom*: speech spectrogram. The example is for the first test case in MuST-C En-De.

System	Similarity
Baseline	0.13
CTC Reg. + translation	0.31
+ CoLaCTC	0.39
CTC Reg. + transcript	0.35
+ CoLaCTC	0.38

Table 5: Cosine similarity of speech representations on the MuST-C En-De test set. We report average results.

System	MT Baseline	CL w/			
		DIV	TRU	LOG	MOD
BLEU	30.5	10.5	15.1	22.9	25.6

Table 6: BLEU scores for *text-to-text* translation conditioned on coarse labels on the MuST-C En-De test set. *MT Baseline*: En→De translation with the vanilla English input; *CL w/ **: using coarse labels instead as the source input. $L = 256$. We use the standard Transformer base setting for experiments.

System	Noun	Verb	Adj.	Adv.	H.Phone
Baseline	43.0	38.6	42.8	46.6	49.4
CTC Reg.					
+ translation	44.4	38.9	44.1	46.8	49.2
+ CoLaCTC	44.3	39.1	43.7	47.5	49.2
CTC Reg.					
+ transcript	45.8	40.6	**46.0**	**48.3**	**51.2**
+ CoLaCTC	**46.1**	**41.1**	45.6	48.0	51.1

Table 7: Translation accuracy of different types of source words on the MuST-C En-De test set. *Adj.*, *Adv.* and *H.Phone* are short for adjective, adverb and homophone.

Coarse labels especially produced by MOD preserve source semantics and are informative for translation. The mappings considered in this study are solely based on heuristic rules. Despite improvements on ST, whether the generated coarse labels themselves are informative is still questionable. We address this concern by performing experiments on text-to-text machine translation and use the coarse label sequence as the source input. Table 6 shows that the coarse labels encode source semantics, achieving non-trivial translation performance. In particular, using MOD achieves a test BLEU score of 25.6. This result still lags far behind the vanilla MT baseline (30.5), but it demonstrates the informativeness of coarse labels, and also partially explains the success of CoLaCTC.

The performance of CoLaCTC is robust on different types of source words. The translation rule for different types of source words often varies greatly. Next, we examine how CoLaCTC generalizes to different words, including nouns, verbs, adjectives, adverbs, and homophones (specific to speech processing). We annotate the part-of-speech tag for each source sentence via Stanford POS tagger (Toutanova et al., 2003), and adopt the homophone list used for contextual evaluation (Zhang
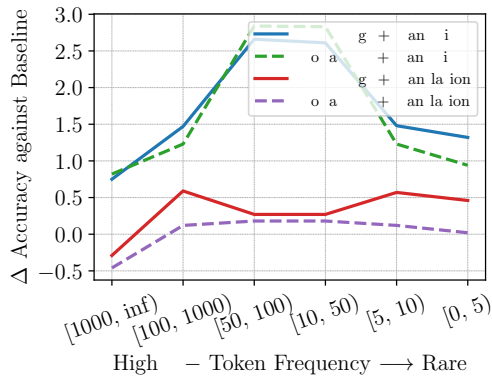
mechanism theoretically is non-trivial. Instead, we understand this question through empirical probes, such as inspecting the change of speech representations and examining how CoLaCTC behaves on different translation aspects.

CTC regularization improves the contextualization of speech representations, so does CoLaCTC. The CTC objective is directly stacked onto the encoder, then what happens to the speech representation (the final encoder output)? Figure 5 illustrates an example, where speech representations after applying CTC regularization (and CoLaCTC) become closer to each other as measured by the cosine similarity. This is further supported by the results in Table 5. Still, the local structure of audio, i.e. the diagonal similarity, is kept. Intuitively, the increased cosine similarity is a reflection of contextualization, and CTC regularization (and CoLaCTC) encourages the encoder to consider (distant) contextual clues.

Figure 6: Relative gains of translation accuracy against the Baseline on MuST-C En-De as a function of token frequency (x-axis). We perform APT for each token group separately.

System	En-It		En-Fr	
	Cov.	Acc.	Cov.	Acc.
Bentivogli et al. (2020)	-	43.3	-	46.0
Baseline	53.3	65.8	59.3	64.6
CTC Reg. + translation	56.0	**67.2**	59.9	66.5
+ CoLaCTC	55.9	67.0	59.6	65.4
CTC Reg. + transcript	55.8	66.9	61.6	66.1
+ CoLaCTC	**57.4**	66.9	**62.0**	**66.7**

Table 8: Coverage (*Cov.*) and accuracy (*Acc.*) scores for gender translation on MuST-SHE En-It and En-Fr.

et al., 2021)[7]. We employ translation accuracy as the metric, approximated by the APT framework (Miculicich Werlen and Popescu-Belis, 2017) where *fast_align* is used to get the word alignment (Dyer et al., 2013). Table 7 shows the results.

CTC regularization largely improves the translation of nouns and adjectives; using transcript as labels further benefits the translation for verbs, adverbs and homophones. Regardless of source word and CTC label types, CoLaCTC shows comparable (sometimes even better) performance to the CTC baseline, showing its strong generalization. Note that observations on other languages, e.g. En-Fr and En-It, are similar (see Appendix B). Further analysis shows that the gains by CTC regularization and CoLaCTC mainly come from benefiting rare-word translation, as shown in Figure 6.

CoLaCTC benefits gender translation similarly to the CTC baseline. Languages often differ in gender expression, leading to translation difficulty. We further evaluate how CoLaCTC handles the gender ambiguity using the MuST-SHE benchmark (Bentivogli et al., 2020). Table 8 shows that CoLaCTC achieves comparable performance

---

[7]The list is publicly available at `bit.ly/3mGITEe`.

to the CTC baseline, suggesting that using coarse labels for CTC regularization doesn't hurt its gender disambiguation ability. Besides, we observe that translation and transcript labels show similar positive effects on gender translation.

7 Conclusion and Future Work

In this paper, we have presented coarse labeling for CTC to address the training inefficiency issue of CTC regularization. The key idea behind CoLaCTC is to transform CTC labels from the vocabulary space to a specified and reduced coarse space. We adopt trivial mappings for this transformation, such as using the modulo operation. Despite its simplicity, CoLaCTC successfully achieves the best of both worlds – improving the training efficiency for CTC regularization (up to $1.77\times$ speedup) and retaining its quality benefits – and generalizes to different types of CTC regularization. Note the training speedup scales as the vocabulary size increases. Our analysis further shows that the genuineness of CTC labels matters less than their origin.

In the future, we are interested in examining the complementarity of CoLaCTC with other advanced ST modeling. We will study how our method performs in a multilingual and simultaneous setup as well as ST settings with extra ASR and/or MT data.

Limitations

While the proposed method achieves encouraging performance across diverse languages and translation setups, our understanding of why it performs so well is still limited, particularly considering the simplicity of the adopted mapping function (Mod). Uncovering the underlying reason behind such success might offer valuable insights to further the speech modeling, having a potential broader impact on the speech processing community.

Acknowledgements

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

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

Xingyu Cai, Jiahong Yuan, Yuchen Bian, Guangxu Xun, Jiaji Huang, and Kenneth Church. 2022. W-CTC: a connectionist temporal classification loss with wild cards. In *International Conference on Learning Representations*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Shun-Po Chuang, Yung-Sung Chuang, Chih-Chiang Chang, and Hung-yi Lee. 2021. Investigating the reordering capability in CTC-based non-autoregressive end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1068–1077, Online. Association for Computational Linguistics.

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.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting Transformer to End-to-End Spoken Language Translation. In *Proc. Interspeech 2019*, pages 1133–1137.

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

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12749–12759.

Yichao Du, Zhirui Zhang, Weizhi Wang, Boxing Chen, Jun Xie, and Tong Xu. 2021. Regularizing end-to-end speech translation with triangular decomposition agreement. *arXiv preprint arXiv:2112.10991*.

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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online. 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.

Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. 2022.

Efficient yet competitive speech translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 177–189, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006a. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Alex Graves, Santiago Fernández, and Faustino Gomez. 2006b. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*, pages 369–376.

Jiatao Gu and Xiang Kong. 2021. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 120–133, Online. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

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.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. 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)*, pages 827–838, Online. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

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

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Zhe Niu and Brian Mak. 2020. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI*, page 172–186, Berlin, Heidelberg. Springer-Verlag.

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, Belgium, Brussels. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108, Online. Association for Computational Linguistics.

Elizabeth Salesky, Matthias Sperber, and Alan W Black. 2019. Exploring phoneme-level speech representations for end-to-end speech translation. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 1835–1841, Florence, Italy. Association for Computational Linguistics.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proc. Interspeech 2021*, pages 3655–3659.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq S2T: Fast speech-to-text modeling with fairseq. 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.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. *arXiv preprint arXiv:2004.10093*.

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.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In *Proceedings of ICML*.

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

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2021. Beyond sentence-level end-to-end speech translation: Context helps. 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 2566–2578, Online. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2021. NeurST: Neural speech translation toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 55–62, Online. Association for Computational Linguistics.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12736–12746. PMLR.

A Experimental Setting

We preprocess texts using Moses scripts (Koehn et al., 2007) and adopt the byte pair encoding algorithm (Sennrich et al., 2016, BPE) to handle rare tokens. In particular, we encode datasets in MuST-C and Multilingual TEDx with a BPE vocabulary size of 8K and 4K, respectively. As for audios, we adopt a sampling rate of 16KHz and filter out segments longer than 3000 frames. We extract 40-dimensional log mel-scale filterbank features for acoustic modeling with a step size of 10ms and window size of 25ms, and further augment them with

	System	Noun	Verb	Adj.	Adv.	H.Phone
	Baseline	48.0	37.5	45.4	44.0	40.9
	CTC Reg. + translation	49.6	38.6	47.2	45.0	41.7
En-It	+ CoLaCTC	49.6	38.7	48.3	45.4	42.0
	CTC Reg. + transcript	50.5	**39.3**	48.0	45.2	42.3
	+ CoLaCTC	**51.2**	39.0	**49.1**	**45.7**	**43.0**
	Baseline	57.2	51.0	53.4	53.9	58.1
	CTC Reg. + translation	58.5	52.8	55.4	54.9	58.9
En-Fr	+ CoLaCTC	58.5	52.3	55.1	54.9	58.2
	CTC Reg. + transcript	59.3	53.0	54.9	55.1	59.7
	+ CoLaCTC	**60.0**	**53.9**	**57.0**	**55.8**	**60.3**

Table 9: Translation accuracy of different types of source words on the MuST-C En-Fr/En-It test set. *Adj.*, *Adv.* and *H.Phone* are short for adjective, adverb and homophone.

their delta and delta-delta features. The final acoustic feature vector is 120-dimensional regularized by mean subtraction and variance normalization.

We focus on bilingual ST and adopt the E2E ST architecture following Zhang et al. (2022): we use Transformer with the post layer normalization structure plus the sinusoidal positional encoding (Vaswani et al., 2017); we set the encoder and decoder depth to 12 and 6, respectively, and adopt the depth-scaled initialization method to stabilize the training (Zhang et al., 2019); we set the model dimension to $d = 256$, the feed-forward layer size to 4096 and the number of attention head to 4; we employ the parameterized distance penalty with $R = 512$ and set $\lambda = 0.3$ for CTC regularization (Zhang et al., 2022).

We train all models via Adam (Kingma and Ba, 2015, $\beta_1 = 0.9, \beta_2 = 0.98$) with a warmup step of 4K and label smoothing rate of 0.1. Samples with around 20K target subwords are scheduled into one batch for training, and we set the maximum training step for MuST-C and Multilingual TEDx to 50K and 20K, respectively. We apply dropout to residual connections and ReLU activations with a rate of 0.2. We perform checkpoint evaluation every 1K training steps on the dev set, and average the best 10 checkpoints for final testing. We use beam search for decoding, and set the beam size to 8. We tune the length penalty for each language pair on its dev set separately.

B   Additional Results and Analysis

The performance of CoLaCTC on different types of source words generalizes to other languages. Table 9 shows the translation accuracy of different models on En-Fr and En-It. The observation is similar to Table 7, where transcript labels are more effective than translation labels for CTC regularization, and that CoLaCTC performs comparable to genuine labels. One exception is that CTC regularization also greatly benefits the translation of verbs and adverbs on En-Fr and En-It. These results suggest that our observation is not language-specific or caused by some random effect, but rather CTC regularization and CoLaCTC generalizes to different language pairs.