Cross-modal Contrastive Learning for Speech Translation

Anonymous ACL submission

Abstract

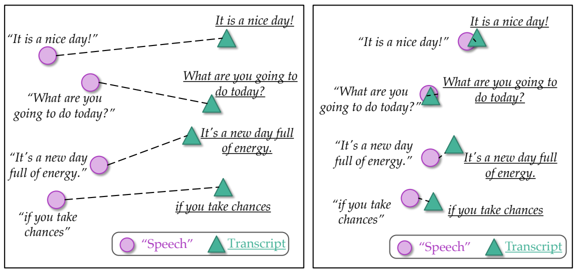
How to learn similar representations for spoken utterances and their written text? We believe a unified and aligned representation of speech and text will lead to improvement in speech translation. To this end, we propose **ConST**, a cross-modal contrastive learning method for end-to-end speech-to-text translation. We evaluate ConST and a variety of previous baselines on multiple language directions (En-De/Fr/Ru) of a popular benchmark MuST-C. Experiments show that the proposed ConST consistently outperforms all previous methods, and achieves the state-of-the-art average BLEU of 28.5. The analysis further verifies that ConST indeed closes the representation gap of different modalities — its learned representation improves the accuracy of cross-modal text retrieval from 4% to 88%.

1 Introduction

End-to-end speech-to-text translation (E2E ST) has been becoming important in many products and real applications. An E2E ST system accepts audio signals as the input and generates the target translation using a single model. Compared with the conventional cascade ST models, E2E ST models have achieved almost comparable (Bentivogli et al., 2021; Dong et al., 2018) or even superior (Ansari et al., 2020; Potapczyk and Przybysz, 2020; Xu et al., 2021) performance.

The performance of an E2E ST model is still restricted by the relatively small parallel data, compared to text machine translation (MT). Existing approaches for ST focus on using additional data from MT and automatic speech recognition (ASR). This can be realized through pre-training approaches (Zheng et al., 2021) or multi-task training frameworks (Tang et al., 2021b; Ye et al., 2021; Han et al., 2021).

Different from the data perspective, this paper investigates the bottleneck of E2E ST from the neural representation perspective. We believe that a right representation for audio input is fundamental to effective speech translation. What is the right representation? A recent neurocognitive study reveals that the human brain processes speech and written text at the same region of the cortex (Regev et al., 2013). Listening to spoken utterance and reading its corresponding sentence result in the same activation patterns in the superior temporal sulcus (Wilson et al., 2018). Drawing an analogy from the human brain to artificial neurons, does this unified representation benefit speech translation?



(a) Current models (b) Expected

Figure 1: Illustration of representations for speech and transcript text (projected to 2D). (a) representations learned by existing models. (b) an ideal representation that we expect, where different modalities with same meaning should stay close to each other.

With this hint from the human brain, we analyze Transformer models for speech translation. We observe a noticeable modality gap between encoder representations of speech and text (Sec. 6 has more details) from existing ST models. An ideal representation should satisfy: if the content of the speech and transcription are similar, their encoded representations should likewise be close to each other. Nevertheless, how to learn unified and aligned speech-text representations?

Inspired by the recent progress of contrastive learning approaches in cross-lingual (Lample and Conneau, 2019; Pan et al., 2021) and cross-modal vision-and-language domains (Li et al., 2021; Zhou et al., 2020; Dong et al., 2019), we designed a simple **con**trastive learning method for **ST** (**ConST**) to learn the representations that meet the aforementioned conditions explicitly. On the one hand, our model inherits the advantages of the previous multi-task learning methods. On the other hand, it reduces the gap between the representations of speech and its corresponding transcription.

Our contributions are as follows.

- We develop ConST for speech translation, a cross-modal contrastive learning method, on top of the multi-task training framework.
- Our experiments on the MuST-C benchmark to show ConST achieves an average BLEU score of 28.5, outperforming the best previous baseline.
- We conduct a cross-modal retrieval experiment and demonstrate that ConST closes the representation gap of two modalities by projecting them into a unified space.

2 Related Work

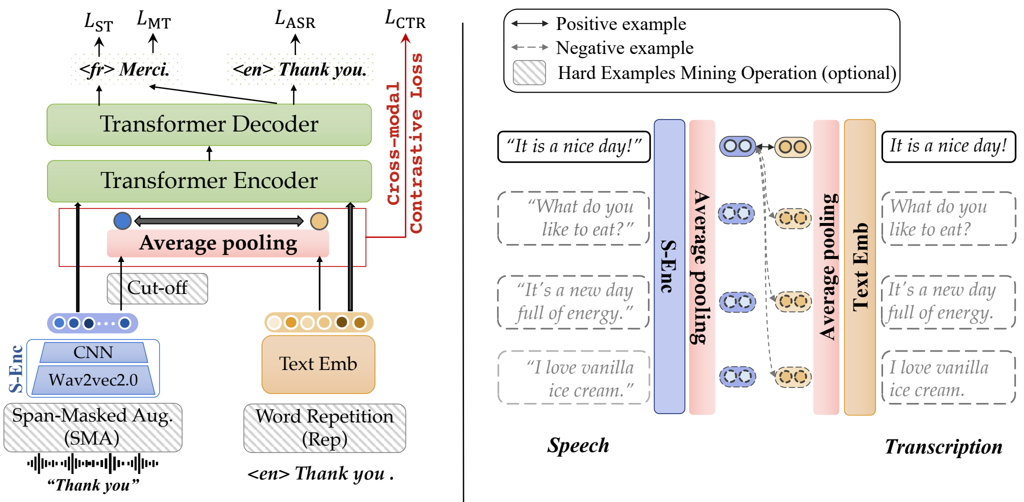
End-to-end ST To alleviate the error propagation in the cascaded ST systems and to make the deployment simpler, Bérard et al. (2016); Weiss et al. (2017) proposed to use an end-to-end architecture to directly translate speech into text in another language, without the intermediate transcription. Kano et al. (2017); Berard et al. (2018); Inaguma et al. (2020); Wang et al. (2020a); Zhao et al. (2021a) implemented several off-the-shelf encoder-decoder E2E-ST models, such as BiLSTM (Greff et al., 2016) and Speech-Transformer (Dong et al., 2018). However, training an end-to-end speech translation model is difficult because we need to design a cross-modal cross-language model, meanwhile, the *speech-transcription-translation* supervised data for speech translation is significantly less than that of MT and ASR. Methods, like data augmentation (Park et al., 2019; Pino et al., 2020; Chen et al., 2021), pre-training (Weiss et al., 2017; Berard et al., 2018; Bansal et al., 2019; Wang et al., 2020b; Alinejad and Sarkar, 2020; Dong et al., 2021a; Zheng et al., 2021), self-training (Pino et al., 2020; Wang et al., 2021), utilizing self-supervised pre-trained audio representation (Wu et al., 2020; Han et al., 2021; Ye et al., 2021; Wang et al., 2021), are proved to be effective. Meanwhile, some work has shown that the encoder-decoder model with a single encoder cannot encode speech information well. For example, Dong et al. (2021b) first proposed a second encoder to further extract semantic information of the speech sequence. Xu et al. (2021) proposed a stacked acoustic-and-textual encoder and introduced large-scale out-of-domain data. Also, multi-task frameworks (Le et al., 2020; Tang et al., 2021b; Ye et al., 2021) are widely applied to further enhance the robustness for ST. As a cross-modal task, some work has noted the problem of the modality gap. (Han et al., 2021) designed a fix-size semantic memory module to bridge such a gap, from the neuroscience perspective. However, we find that this approach actually sacrifices the effect of MT. So in this paper, we propose a simple yet effective contrastive learning method to bridge the gap and to improve ST performance.

Contrastive learning Our method is motivated by the recent success in contrastive representation learning. The contrastive learning method was first proposed to learn representations from unlabeled datasets (hence the term, self-supervised learning) by telling which data points are similar or distinct, especially in the field of computer vision (Chopra et al., 2005; Gutmann and Hyvärinen, 2010; Schroff et al., 2015; Sohn, 2016; Oord et al., 2018). Khosla et al. (2020) extended the self-supervised batch contrastive approach to the fully-supervised setting and proposed a supervised contrastive learning method. In speech processing, representative methods focused on speaker identification (Ravanelli and Bengio, 2018), speech recognition (Schneider et al., 2019), and audio representation learning (van den Oord et al., 2018; Baevski et al., 2020). In the NLP area, the contrastive framework is used for sentence representation learning (Fang and Xie, 2020; Shen et al., 2020; Gao et al., 2021; Wu et al., 2021; Yan et al., 2021) and machine translation Pan et al. (2021). Very recently, contrastive learning is also applied to learning a unified representation of image and text (Dong et al., 2019; Zhou et al., 2020; Li et al., 2021). Motivated by the contrastive learning frameworks in cross-lingual and cross-modal topics, we introduce a similar idea in speech translation.

3 The ConST Approach

An end-to-end speech translation model directly translates audio sequence $\mathbf{s} = (s_1, ..., s_{|\mathbf{s}|})$ to the text $\mathbf{y} = (y_1, ..., y_{|\mathbf{y}|})$ in the target language. Speech translation corpus $\mathcal{D} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$ provides *transcript* $\mathbf{x} = (x_1, ..., x_{|\mathbf{x}|})$ in the source language, as well.

In this section, we present the overall speech translation model and cross-modal contrastive learning. We also provide several feasible strategies to construct more positive and negative pairs



(a) Overall Model Structure (b) Cross-modal Contrastive Learning

Figure 2: **Left**: Model structure of ConST. The gray shaded modules are the *optional* data augmentation operations introduced in Section 3.3. **Right**: An illustration of cross-modal contrastive learning.

to enhance the contrastive learning.

3.1 Model Framework

Our model consists fout sub-modules: a speech encoder, a word embedding layer, a Transformer Encoder and a Transformer decoder (Figure 2). It is designed to take either speech or a sentence as input, and to output either source transcript or target translation text. Such architecture enables a universal framework for multiple tasks, including ST, MT and ASR.

The *speech encoder* module (S-Enc) is designed to extract low-level features for speech signals. It contains Wav2vec2.0 (Baevski et al., 2020) and two additional convolutional layers. The input is raw waveform signal sampled at 16kHz. Each convolutional layer has a stride of 4 and d channels. In total, it shrinks the time dimension by a factor of 4. Denote $\mathbf{a} = \text{S-Enc}(\mathbf{s})$ as the audio representation of the speech, $|\mathbf{a}| \ll |\mathbf{s}|$.

Parallel to the speech encoder is the *word embeeding layer*. It is the same as word embedding for text translation.

Both the speech encoder and word embedding layer are connect to *Transformer encoder* and then passed to the *Transformer decoder*. The Transformer encoder and decoder are using the same configuration as the original (Vaswani et al., 2017). To explain, the *Transformer encoder* further extracts the high-level semantic hidden representation of two modalities. The *Transformer decoder* generates the word sequences (transcription and translation) for ST, MT and ASR tasks. Since our model has a complete Transformer encoder-decoder as a sub-module, this makes it possible to pre-train using large-scale extra MT parallel data.

Previous work has shown that multi-task learning on ST, MT and ASR improves translation performance (Indurthi et al., 2020; Tang et al., 2021b; Ye et al., 2021). Our training loss consists of the following elements.

$$\mathcal{L} = \mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{ASR}} + \mathcal{L}_{\text{MT}} + \lambda \mathcal{L}_{\text{CTR}} \quad (1)$$

where

$$\mathcal{L}_{\text{ST}} = -\sum_n \log P(\mathbf{y}_n|\mathbf{s}_n)$$

$$\mathcal{L}_{\text{ASR}} = -\sum_n \log P(\mathbf{x}_n|\mathbf{s}_n)$$

$$\mathcal{L}_{\text{MT}} = -\sum_n \log P(\mathbf{y}_n|\mathbf{x}_n)$$

The first three elements are cross-entropy losses on <*speech, target text*>, <*speech, source text*> and <*source text, target text*> pairs. These pairs are built from the triplet ST data. We also introduce a cross-modal contrastive loss term $\mathcal{L}_{\text{CTR}}$ (see Section 3.2 for details). It aims to bring the representation between the speech and textual transcription modalities closer (its effect will be analyzed in detail in Section 6). λ is a tuned hyper-parameter of the weighted contrastive loss term.

3.2 Cross-modal Contrastive Learning

As mentioned in the beginning, since we need to produce similar representations for the speech and transcript sharing the same semantic meanings, we propose cross-modal contrastive learning method to bring their representations closer together. The main idea of cross-modal contrastive learning is to introduce a loss that brings speech and its corresponding transcript (positive example) near together while pushing irrelevant ones (negative examples) far apart.

Given a positive example of such a speech-transcript pair $(\mathbf{s}, \mathbf{x})$, we randomly pick a set of $N-1$ transcripts $\{\mathbf{x}_i^-\}_{i=1}^{N-1}$ from the same batch as negative examples. For speech $\mathbf{s}$ and its transcript $\mathbf{x}$, we first average them in terms of the time dimension,

$$u = \text{MeanPool}(\text{S-Enc}(\mathbf{s})) \tag{2}$$

$$v = \text{MeanPool}(\text{Emb}(\mathbf{x})) \tag{3}$$

and apply the multi-class N-pair contrastive loss (Sohn, 2016):

$$\mathcal{L}_{\text{CTR}} = -\sum_{s,x} \log \frac{\exp(sim(u,v)/\tau)}{\sum_{x_j \in \mathcal{A}} \exp(sim(u, v(x_j))/\tau)} \tag{4}$$

where $\mathcal{A} = \{\mathbf{x}\} \cup \{\mathbf{x}_i^-\}_{i=1}^{N-1}$, τ is the temperature hyper-parameter, and sim is the cosine similarity function $sim(a,b) = a^\top b/\|a\|\|b\|$. In the implementation, negative examples $\{\mathbf{x}_i^-\}_{i=1}^{N-1}$ are from the same training batch of data (Figure 2(b)).

3.3 Mining Hard Examples for Contrastive Learning

To further enhance the contrastive learning, we introduce three strategies to mine additional hard examples. These strategies are at input and representation (gray shaded modules in Figure 2(a)). Specific schematic illustrations of each operations are shown in Figure 3.

Span-Masked Augmentation We mask consecutive segments of an original audio waveform sequence $\mathbf{s}$ to obtain a new modified speech $\mathbf{s}'$. We take $\mathbf{s}'$ as an input to the model, and compute the contrastive loss its original corresponding transcript. We randomly sample without replacement all time steps in the original waveform of the speech to be the starting indices with a probability p, and then we set the sub-sequence M successive time steps to be blank. In the experiment, we tried multiple configurations, and found $p = 0.25, M = 3600$ the best, resulting in a masked span of 0.225 second. Since the masked speech fragment is very short, we consider the masked speech and the original transcript to be positive pairs, and the remaining transcripts in the same batch to be negative pairs.

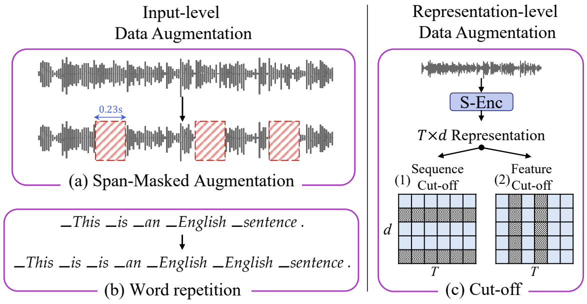


Figure 3: Schematic illustration of the data augmentation strategies. In the cut-off strategy, the gray shaded grid represents the zero-out element.

Word Repetition The word repetition strategy randomly replicates some words (or sub-words) in the original sentences, with two advantages for improving representation robustness. First, as the length of the sentence is shorter than that of its audio representation, randomly repeating the words in the sentence is a simple yet useful technique to increase the length. Second, repeating words does not change the semantics and is suitable as an extra positive example of the corresponding speech. Specifically, given sentence $\mathbf{x}$, each sub-word token x_i can be duplicated k **more** times, resulting in the duplicated sentence $\mathbf{x}'$, where $k = 0, 1, 2, ...$ and $k \sim \text{Poisson}(1)$. We regard $\mathbf{x}'$ as the additional positive example for the speech $\mathbf{s}$ and the samples with the same operation in the same batch as the negative examples.

Cut-off strategy Recent studies on natural language understanding and generation have proved cut-off data augmentation strategy to be successful (Shen et al., 2020; Yan et al., 2021). We analogize a similar idea to the cut-off data augment approach for speech representation. We entirely erase the $T \times d$ representation matrix along each dimension and set the erased terms to 0. Here, we present two variants: ***sequence cut-off***, which erases some sequence dimension, and ***feature cut-off***, which erases some feature dimension. Note that there is a difference between cut-off and dropout. Dropout randomly sets some elements to 0, while cut-off is a dimensional “block" dropout. Similarly, we treat the cut-off audio representation and the original transcribed sentence as positive pairs, and the rest sentences in the same batch as negative pairs.

4 Experiments

4.1 Experimental Setups

ST datasets We conduct experiments on three representative directions from **MuST-C**

dataset [1] (Di Gangi et al., 2019): En-De, En-Fr and En-Ru. Due to the computation limatation, we do not preform for the rest language directions. As one of the largest ST benchmarks, MuST-C contains more than 385 hours of TED talks for each direction.

MT datasets We introduce external **WMT** datasets (Bojar et al., 2016) for each translation direction, as the expanded setup.

Table 6 (in Appendix. A) lists the statistics of all the datasets included.

Model Configurations The Wav2vec2.0 in the S-Enc is only pre-trained on Librispeech (Panayotov et al., 2015) speech only without any downstream fine-tuning[2]. Two layers of CNNs after the Wav2vec2.0 are set to kernel size 5, stride size 2 and hidden size 512. The Transformer follows the base configuration, with 6 layers of encoder and decoder, hidden size $d = 512$, 8 attention heads, and 2048 FFN hidden states. We use pre-layer normalization for stable training.

Experiment Details We evaluate case-sensitive detokenized BLEU using sacreBLEU[3] (Post, 2018) on MuST-C `tst-COMMON` set. We also report the ChrF++ score [4] (Popović, 2017) and translation error rate (TER) [5] in the analysis. We use the raw 16-bit 16kHz mono-channel *speech* input. We jointly tokenize the bilingual *text* using SentencePiece (Kudo and Richardson, 2018), with a vocabulary size of 10k. For the training loss, we set contrastive temperature $\tau = 0.02$ and weight of contrastive term $\lambda = 1.5$.

Appendix B contains more detailed settings and explanations for the baseline models in Table 1. Appendix C shows the experiments on the choice of the hyper-parameters.

4.2 Main Results

Comparison with end-to-end ST models Table 1 shows the main results. Since many existing works regard "leveraging external data" to be one of their model's features, their strong performances are largely predicated on the utilization of auxiliary data, especially large-scale MT data. For a relatively fair comparison, we investigate two cases: (1) without external MT data and (2) with external MT data. Without the external MT data, our method already gains an average improvement of 0.5 BLEU over the previous best models. Also when speech data is introduced for pre-training, our method works better than others (Self-training, W-Transf. and XSTNet). When extra MT data are introduced, our method also outperforms SOTA by an average of 0.7 BLEU. Among the benchmark models, with the same goal of closing two modality gaps, Chimera (Han et al., 2021) constructed an extra fixed-length shared semantic space. However, the shared memory with fixed size actually compromises the MT performance, while our contrastive learning approach is more straightforward and effective.

Comparison with cascaded ST systems We compare our method with several cascade baselines, where Ye et al. (2021) and Xu et al. (2021) provided two strong cascade systems trained using MuST-C and external ASR and MT data (LibriSpeech, WMT, and Opensubtitles). From Table 2, we find that as an end-to-end model, ConST can outperform these strong cascade models. In Appendix E, we provide a case study to show such improvement.

5 Analysis

5.1 Is contrastive loss effective?

With the same model architecture and the same pre-training + fine-tuning procedure, the main difference between ConST and XSTNet (Ye et al., 2021) is whether we use the contrastive loss term during the fine-tuning or not. Comparing the BLEU results of w/o and w/ external MT data situations in Table 1, we find that ConST further improves 0.5 and 0.7 BLEU scores in terms of three translation directions on average. This demonstrates the effectiveness of the cross-modal contrastive learning.

5.2 Which layer to contrast on?

An intriguing question is which representations should be considered in the contrastive loss function. In the method part (Section 3.2), we use averaged audio representation u for speech $\mathbf{s}$ (Eq.(2)) and averaged lexical embedding v for the transcript $\mathbf{x}$ (Eq.(3)), denoted as *low-level repr.*. Whereas inspired by a recent study in multilingual MT (Pan et al., 2021), we also provide an alternative contrastive loss as a comparison, whose speech and

[1] We use v1.0. https://ict.fbk.eu/must-c/

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

[3] https://github.com/mjpost/sacrebleu, **BLEU Signature**: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0

[4] **ChrF2++ Signature**: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:yes | nc:6 | nw:2 | space:no | version:2.0.0

[5] **TER Signature**: nrefs:1 | bs:1000 | seed:12345 | case:lc | tok:tercom | norm:no | punct:yes | asian:no | version:2.0.0

Models	External Data				BLEU			
	Speech	Text	ASR	MT	En-De	En-Fr	En-Ru	Avg.
w/o external MT data								
Fairseq ST (Wang et al., 2020a)	-	-	-	-	22.7	32.9	15.3	23.6
NeurST (Zhao et al., 2021a)	-	-	-	-	22.8	33.3	15.1	23.7
Espnet ST (Inaguma et al., 2020)	-	-	-	-	22.9	32.8	15.6	23.8
Dual Decoder (Le et al., 2020)	-	-	-	-	23.6	33.5	15.2	24.1
W-Transf. (Ye et al., 2021)	✓	-	-	-	23.6	34.6	14.4	24.2
Speechformer (Papi et al., 2021)	-	-	-	-	23.6	-	-	-
Self-training (Pino et al., 2020)	✓	-	✓	-	25.2	34.5	-	-
SATE (Xu et al., 2021)	-	-	-	-	25.2	-	-	-
BiKD (Inaguma et al., 2021)	-	-	-	-	25.3	35.3	-	-
XSTNet (Ye et al., 2021)	✓	-	-	-	25.5	36.0	16.9	26.1
Mutual-learning (Zhao et al., 2021b)	-	-	-	-	-	36.3	-	-
ConST	✓	-	-	-	**25.7**	**36.8**	**17.3**	**26.6**
w/ external MT data								
MTL (Tang et al., 2021b)	-	-	-	✓	23.9	33.1	-	-
LightweightAdaptor (Le et al., 2021)	✓	✓	-	✓	24.6	35.0	16.4	25.3
FAT-ST (Big) (Zheng et al., 2021)	✓	✓	✓	✓	25.5	-	-	-
JT-S-MT (Tang et al., 2021a)	-	-	-	✓	26.8	37.4	-	-
Chimera (Han et al., 2021)	✓	-	-	✓	27.1†	35.6	17.4	26.7
XSTNet (Ye et al., 2021)	✓	-	-	✓	27.1	38.0	18.4	27.8
SATE (Xu et al., 2021)	-	-	✓	✓	28.1†	-	-	-
ConST	✓	-	-	✓	**28.3**	**38.3**	**18.9**	**28.5**

Table 1: Case-sensitive detokenized BLEU scores on MuST-C `tst-COMMON` set. "Speech" denotes unlabeled speech data. "Text" means unlabeled text data, *e.g.* Europarl V7 (Koehn et al., 2005), CC25 (Liu et al., 2020a). † use external 40M OpenSubtitles (Lison and Tiedemann, 2016) MT data. Other models only use WMT data.

Models	En-De	En-Fr	En-Ru
Cascaded			
Espnet(Inaguma et al., 2020)	23.6	33.8	16.4
(Ye et al., 2021)	25.2	34.9	17.0
(Xu et al., 2021)	28.1	-	-
End-to-end			
ConST	**28.3**	**38.3**	**18.9**

Table 2: ConST versus the cascaded ST systems on MuST-C En-De/Fr/Ru test sets. Ye et al. (2021) and Xu et al. (2021) are two strong cascaded models.

Representations	BLEU	ChrF++	TER
low-level repr.	**28.3***	**53.2***	**59.4***
high-level repr.	27.5†	52.6†	61.0
w/o contrative loss	27.1	52.1	61.0

Table 3: BLEU, ChrF++ and TER (%) on En-De test set. Different representations are tested. *: ConST is significantly better than the other two baselines ($p < 0.01$). †: the model is significantly better the baseline model without contrastive loss ($p < 0.05$).

text features are average-pooled semantic representations derived from the Transformer encoder, denoted as *high-level repr.*.

Table 3 shows that contrastive learning using **the low-level representations (Line 1) is better than using the high-level ones (Line 2)**. On the other hand, although the performance of Line 2 is relatively inferior, it still outperforms the multi-task model without the contrastive loss (Line 3).

5.3 Is contrastive loss better than other losses?

Our goal for introducing the contrastive loss term (denoted as CTR Loss) is to close the distance between speech and text representations. Whereas, there are other options to achieve this goal, such as L2 loss and CTC loss.

- **L2 Loss**: Without introducing any negative samples, L2 loss directly reduces the Euclidean distance between the representations of two modalities by minimizing $\mathcal{L} = \|u - v\|^2$. L2 loss can be viewed as an implementation based on the idea of knowledge distillation (Heo et al., 2019; Dong et al., 2021b).
- **CTC Loss**: The connectionist temporal classification (CTC) loss (Graves et al., 2006) is commonly used in speech-related tasks (Xu et al., 2021; Dong et al., 2021b). Unlike contrastive loss that cares about the representation, CTC loss connects the two modalities by establishing speech-text alignment and maximizing $p(\mathbf{x}|\mathbf{a}) =$

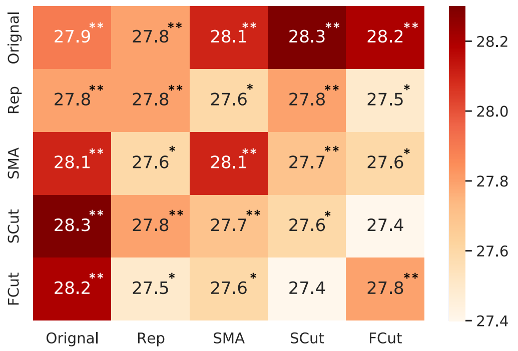
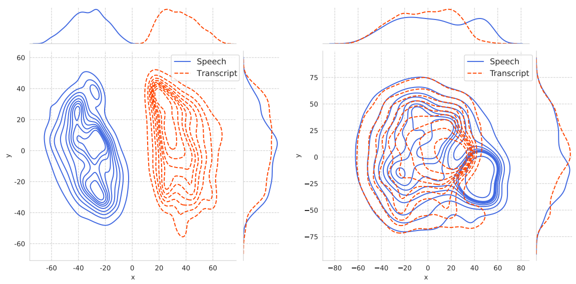


Figure 4: The heat map visualization of the BLEU scores on En-De test set, with 5×5 combinations of the original contrastive loss (Original) and data augmentation methods – word repetition (Rep), span-masked augmentation (SMA), sequence cut-off (SCut) and feature cut-off (FCut). * and ** mean the improvements over the baseline without contrastive loss are statistically significant (*:$p < 0.05$, **:$p < 0.01$).



(a) w/o CTR loss (b) ConST

Figure 5: Bivariate KDE contour plot of the representation of speech and transcript in source language English. T-SNE is used to reduce into 2D. The blue lines are the audio representations and the red dashed lines stand for text. (a) for the vanilla multi-task framework without any extra supervision. (b) for our proposed ConST model. Sentences are from En-De test set.

$\sum_{\pi\in\Pi_{\mathbf{s},\mathbf{a}}}\prod_{t=1}^{T} p_t(\pi_t|\mathbf{a})$, where $\Pi_{\mathbf{s},\mathbf{a}}$ is the set of all valid alignments.

Compared to the other two ways of bridging the modality gap, L2 and CTC loss, is the contrastive loss term better? The answer is yes according to the results in Table 4. Our explanation is that information on the negative samples benefits the contrastive loss, bringing the the distance between the speech and its corresponding transcription closer while pushing the distance to the irrelevant text farther.

Extra Loss	BLEU	ChrF++	TER
CTR Loss	**28.3***	**53.2**	**59.4***
CTC Loss	27.5†	53.0†	60.1†
L2 Loss	27.3	52.4	60.7
-	27.1	52.1	61.0

Table 4: BLEU, ChrF++ and TER (%) on En-De test set under different loss terms other than the basic multi-task NLL loss. *: ConST is significantly ($p < 0.01$) better than the other three alternatives. †: the improvement from CTC loss over the baseline without extra loss is significant ($p < 0.01$).

5.4 Analysis on the data augmentation strategies

In Section 3.3, we proposed four methods to mine the hard examples for contrastive learning, namely span-masked augmentation (SMA), word repetition (Rep), sequence cut-off (SCut), and feature cut-off (FCut). In this section, we study how effective these methods are, and to do so, we consider the BLEU performances of their 5×5 combinations (Figure 4). Note that "Original" means the original contrastive loss in Eq.(4) without any data augmentation, and the diagonal in the heat map represents only one strategy used. For an easy and fair comparison, we set the weight of the contrastive term to 1.0 uniformly. We have the following observations.

All the data augmentation methods are effective. All the BLEU scores in Figure 4 exceed the strong multi-task model trained without contrastive learning (27.1). Among all the strategies, the combination of the original and SCut reaches the best result (28.3), and is better than the model without any expanded operations ($p < 0.01$). Generally, to find the best model, we suggest adopting multiple strategies and choosing the best checkpoint on the dev-set.

The combinations of the data augmentation methods and the "original" have relatively better performances. We argue that we need the original positive and negative examples to give more accurate representations (without any dropout) for contrastive learning. On the contrary, without the help of "original" loss, the performance with both sequence cut-off and feature cut-off is the worst in Figure 4, probably because too much information is lost by superimposing the two.

6 Why does cross-modal contrastive learning work? — Analysis on the Modality Gap

As mentioned earlier, the existing multi-task training models cannot address the *speech-text modality gap*. Does ConST reduce the representation gap between speech and text?

6.1 Visualization of Representation

Does the *speech-text modality gap* exist without explicitly bridging the two? *Speech-text modality gap* means the discrepancy between the audio representations and transcription sentence embeddings. To visualize it, we plot the bivariate kernel density estimation (Parzen, 1962) (KDE) contour of the dim-reduced feature of them, where T-SNE (Van der Maaten and Hinton, 2008) is used to reduce the dimension into two (Figure 5). Ideally, if the representations of speech and its corresponding transcript are similar, their KDEs will be similar, and thus the contour lines will overlap as much as possible. However, Figure 5(a) is the KDE contour of the multi-task framework without any explicit modeling to bring two modalities together (Ye et al., 2021). It shows that the representations are so dissimilar that they are organically divided into two clusters, *i.e. speech-text modality gap* exists.

Does ConST reduce the modality gap? As shown in Figure 5(b), compared to the baseline model without contrastive learning, ConST with cross-modal contrastive learning is able to bring representations of different modalities much closer. This means that the audio representation contains more linguistic information similar to that of the textual transcription, which is more advantageous for the downstream ST generation through the shared *Transformer encoder* and *Transformer decoder*.

6.2 Cross-modal Retrieval

How good is the cross-modal representation space learned from ConST? To answer this question, we conduct a retrieval experiment, *i.e.* finding the nearest (smallest cosine similarity) transcript based on the speech representation. We compare ConST model with the baseline without cross-modal contrastive learning and report the top-1 retrieval accuracy using (1) the low-level representations and (2) the high-level semantic representations, in Table 5.

When retrieving the text using low-level representations, our method gains a substantial 79% increase over the baseline. In addition, we find that without explicit contrastive modeling, the baseline can achieve retrieval accuracy up to 94% according to the semantic representations outputted from the *Transformer encoder*. We believe that such high accuracy is automatically learned from the triple-supervised data itself under the multi-task learning framework. With such a degree of cross-modal alignment, if we construct the contrastive loss with semantic representations, its gain to the ST performance turns out to be limited, which exactly corroborates the findings in Section 5.2 – low-level representations are preferred in the cross-modal contrastive learning.

Representations	CTR loss	Acc.
low-level repr.	×	9.4
	✓	**88.6**
high-level repr.	×	94.7
	✓	**95.0**

Table 5: Cross-modal top-1 retrieval accuracy (%) on En-De test set. Two different representations are used, based on which, ConST achieves huge accuracy improvements.

7 Conclusion

In this paper, we propose ConST, a simple yet effective contrastive learning framework bridging the speech-text representation gap and facilitating the ST with limited data. We also provide feasible data augmentation methods to learn robust representations. The results on the MuST-C ST dataset prove the effectiveness of the method.

8 Broader Impact

This work improves the performance of ST tasks on public datasets by learning speech representations that are more similar to text representations, but the model is far from being achieved for industrial-grade implementations. In real scenarios, for example, the original voice is noisier and the distribution of speech lengths is more complex than in the public dataset, which cannot be handled by an end-to-end model alone. The shortcoming of this model is that it still needs a certain amount of labeled data for training, especially <*speech,transcription*> to learn better speech representation, and for the more than $7,000$ languages and dialects in the world, most of them do not have corresponding translations or even transcriptions, our method does not work in untranscribed scenarios. In this paper, we focus on the improvement brought by the better speech representation on the ST task, and obtained good results with hundreds of hours of speech data. We hope that our work achieves better results using more data (*e.g.* raw speech, raw text, ASR, MT data) in the future.

References

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proc. of EMNLP*, pages 8014–8020.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, et al. 2020. Findings of the iwslt 2020 evaluation campaign. In *Proc. of IWSLT*, pages 1–34.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. of NeurIPS*.

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

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proc. of ACL*.

Alexandre Berard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *Proc. of ICASSP*, pages 6224–6228.

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*.

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

Junkun Chen, Mingbo Ma, Renjie Zheng, and Liang Huang. 2021. Specrec: An alternative solution for improving end-to-end speech-to-text translation via spectrogram reconstruction. *Proc. of InterSpeech*, pages 2232–2236.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proc. of CVPR*, volume 1, pages 539–546. IEEE.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. of NAACL-HLT*, pages 2012–2017.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proc. of NeurIPS*, pages 13063–13075.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *Proc. of ICASSP*, pages 5884–5888.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021a. Consecutive decoding for speech-to-text translation. In *Proc. of AAAI*.

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

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *CoRR*, abs/2005.12766.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proc. of EMNLP*.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of ICML*, volume 148, pages 369–376.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

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

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Proc. of ACL - Findings*.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. 2019. A comprehensive overhaul of feature distillation. In *Proc. of the ICCV*, pages 1921–1930.

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

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 *Proc. of ACL*, pages 302–311.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. Data efficient direct speech-to-text translation with modality agnostic meta-learning. In *Proc. of ICASSP*. IEEE.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Proc. of INTERSPEECH*, pages 2630–2634.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Proc. of NeurIPS*, 33.

Philipp Koehn et al. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of EMNLP*, pages 66–71.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proc. of NeurIPS*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proc. of COLING*, pages 3520–3533.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proc. of ACL*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proc. of LREC*, pages 923–929.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL - Demonstrations*, pages 48–53.

Xiao Pan, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proc. of ACL*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proc. of ICASSP*, pages 5206–5210.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Speechformer: Reducing information loss in direct speech translation. In *Proc. of EMNLP*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. 2019. Specaugment: A simple augmentation method for automatic speech recognition. In *Proc. of INTERSPEECH*.

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

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. In *Proc. of INTERSPEECH*, pages 1476–1480.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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.

Tomasz Potapczyk and Paweł Przybysz. 2020. Srpol's system for the iwslt 2020 end-to-end speech translation task. In *Proc. of IWSLT*, pages 89–94.

Mirco Ravanelli and Yoshua Bengio. 2018. Learning speaker representations with mutual information. *arXiv preprint arXiv:1812.00271*.

Mor Regev, Christopher J Honey, Erez Simony, and Uri Hasson. 2013. Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, 33(40):15978–15988.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. of INTERSPEECH*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. of CVPR*, pages 815–823.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proc. of NeurIPS*, volume 29, pages 1857–1865. Curran Associates, Inc.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021a. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proc. of ACL*.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *Proc. of ICASSP*, pages 6209–6213. IEEE.

Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *Proc. of ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3915–3923.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proc. of AACL*, pages 33–39.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021. Large-scale self-and semi-supervised learning for speech translation. *arXiv preprint arXiv:2104.06678*.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. In *Proc. of ACL*, pages 3728–3738.

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

Stephen M. Wilson, Alexa Bautista, and Angelica McCarron. 2018. Convergence of spoken and written language processing in the superior temporal sulcus. *NeuroImage*, 171:62–74.

Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. In *Proc. of INTERSPEECH*.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *arXiv preprint arXiv:2109.04380*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proc. of ACL*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2021a. NeurST: Neural speech translation toolkit. In *Proc. of ACL - System Demonstrations*.

Jiawei Zhao, Wei Luo, Boxing Chen, and Andrew Gilman. 2021b. Mutual-learning improves end-to-end speech translation. In *Proc. of the EMNLP*, pages 3989–3994.

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

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proc. of AAAI*, volume 34, pages 13041–13049.

A Statistics of all datasets

En→	ST (MuST-C) hours	ST (MuST-C) #sents	MT name	MT #sents
De	408	234K	WMT16	4.6M
Fr	492	292K	WMT14	40.8M
Ru	489	270K	WMT16	2.5M

Table 6: Statistics of all datasets

B Experimental Details

Training and Implementation Details We use Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.98$) with learning rate $= 1e^{-4}$ and warmup 25k steps during the ST training. We also implement the expanded setting with the introduction of external WMT to train the Transformer module. In the pre-training stage, we set the learning rate $= 7e^{-4}$ and warmup 4000 steps. For robust training, we set label smoothing to 0.1, and dropout rate to 0.1. The hyper-parameters for different data augmentation methods are as follows: for masked audio span strategy, we set masking probability $p = 0.25$ and masking span length $M = 3600$ frames; for both sequence and feature cut-off, we set the cut-off dropout rate as 0.1. We save the checkpoint with the best BLEU on dev-set and average the last 10 checkpoints. For decoding, we use a beam size of 10 and length penalty 0.7 for German, 1.0 for French, and 0.4 for Russian. We train the models in 8 Nvidia Tesla V100 GPUs for each experiment. We use Fairseq (Ott et al., 2019) as the code-base for our implementation.

Baseline Models In Table 1, we compared our method with end-to-end baseline models whose audio inputs are 80-channel log Mel-filter bank, including: FairseqST (Wang et al., 2020a), NeurST (Zhao et al., 2021a), Espnet ST (Inaguma et al., 2020), Dual-decoder Transformer (Le et al., 2020), SATE (Xu et al., 2021), Speechformer (Papi et al., 2021), self training (Pino et al., 2020) and mutual learning (Zhao et al., 2021b) method, STAST (Liu et al., 2020b), bi-KD (Inaguma et al., 2021), MLT method (Tang et al., 2021b), Lightweight Adaptor (Le et al., 2021), and JT-S-MT (Tang et al., 2021a), FAT-ST (Zheng et al., 2021), We also compare our method to baseline models that have pretrained Wav2vec2.0 as a module, including:

- **W-Transf.** (Ye et al., 2021): the model has the same structure as ours, but is only trained on <*speech, translation*> parallel data.
- **Chimera-ST** (Han et al., 2021): the model that builds a shared semantic memory for both audio and text modalities.
- **XSTNet** (Ye et al., 2021): the model has the same structure as ours, and adopted a multi-task fine-tuning strategy.

C The Choice for Hyper-parameters

Influence of Temperature In the contrastive loss, the temperature hyper-parameter is provided to control the smoothness of the distribution normalized by softmax operation. A high temperature helps to smooth the distribution, making it more difficult for the model to distinguish between positive and negative samples (corresponding to correct transcriptions and other transcriptions in this work), while the low temperature behaves just the opposite. We choose several temperature hyper-parameters ranging from 0.01 to 0.5, and Figure 6 shows their BLEUs on the test and dev sets . We find that (1) the choice of the temperature does not drastically affect the final BLEU score, and (2) we recommend that the temperature τ be set between 0.02 and 0.05 to ensure a relatively good ST performance. In the experiment, we use $\tau = 0.02$.

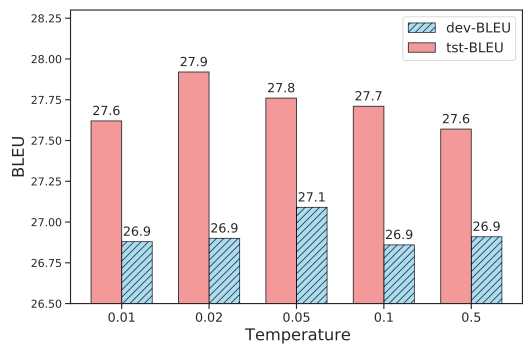


Figure 6: En-De BLEU scores on `tst-COMMON` and `Dev` set. the x-axis is the choices of different temperature τ in Eq.(4) varying from 0.01 to 0.5.

Influence of Contrastive Loss Weight The total loss we optimize, Eq.(1), is a linear combination of the multi-task cross-entropy losses $\mathcal{L}_{\text{MLT}}$ and the contrastive term $\mathcal{L}_{\text{CTR}}$. To investigate how much the contrastive terms affect BLEU, we fix its temperature $\tau = 0.02$, adjust the values of its loss weight λ from 0.1 to 2.0, performed three experiments for each value, and test the average BLEU on En-De `tst-COMMON` set. Figure 7 depicts the performances. First, all objective functions containing

$\mathcal{L}_{\text{CTR}}$, even if their weights λ take different values, are apparently better than the baseline model with $\mathcal{L}_{\text{MLT}}$ only $\mathcal{L}_{\text{CTR}}$. Then, the best BLEU score is achieved at loss weight $\lambda = 1.5$, corresponding to the results in Table 1. And when analyzing the effect of data augmentation strategies (Section 5.4), since we need to consider the combination between them, which is more complicated. Therefore, we set the loss weight to 1.0 uniformly for simplicity. In general, we recommend that the weight hyper-parameter takes a value between 0.8 and 1.5.

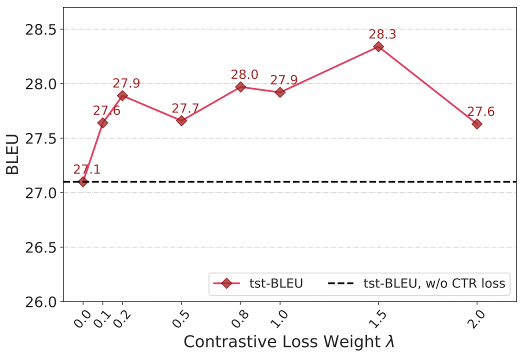


Figure 7: En-De BLEU scores on `tst-COMMON` and `Dev` sets. The x-axis is the weight of the contrastive loss term λ in Eq.(1). Experiments are performed under the fix temperature hyper-parameter $\tau = 0.02$.

D Data Scale for Fine-tuning

The experiments in the main paper show that our model can perform well by introducing external MT data pre-training. Here, we simulate the scenario with plenty of MT and speech data and limited ST triple-labeled data, and does ConST have the ability of low-resource learning? In the experiment, we reduce the labeled ST data to 1, 10, and 100 hours, corresponding to sentence counts of about 500, 5k, and 50k sentences. For a fair comparison, we use the same MT pre-trained Transformer module as in the main paper. We find the contrastive loss particularly helpful when the amount of speech data is extremely small, like only 1 hour of speech. Second, the multi-task training strategy is also very effective in improving the robustness of the model performance. We also find that by using easily accessible MT and speech pre-training, our model could reach the previous baseline results without pre-training using only $1/4$ of the original data, *i.e.* 100 hours of labeled ST data.

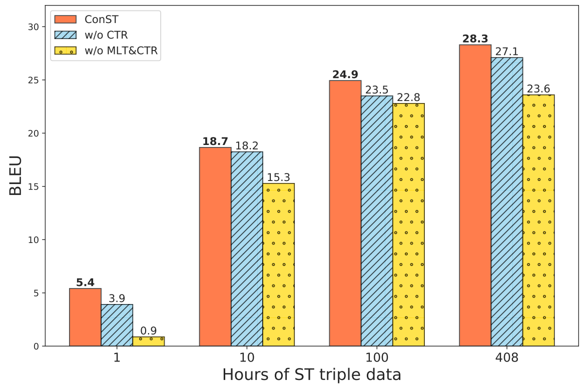


Figure 8: En-De BLEU scores on `tst-COMMON` sets. The horizontal axis is the amount of ST data (in hours of speech).

E Case Analysis

In this section, we use several cased that our proposed ConST model generates to compare our model with the cascaded model and the previous end-to-end model, XSTNet[6] (Ye et al., 2021).

For this first case, the cascaded system fails to give a right translation due to the mis-punctuation issue (*klingt* is a verb), while the end-to-end model, XSTNet and ConST translate correctly. For the second case, the previous end-to-end XSTNet model cannot accurately translate the phrase "started exploring this idea of", which performs worse than the cascaded one. Whereas ConST successfully conveys the meaning of "this idea" , and translates more accurately than XSTNet. We believe this improvement comes from the cross-modal contrastive learning.

[6]The generation cases of the previous models can be found at `https://reneeye.github.io/projects/XSTNet`.

Models		
		CASE 1
Ref.	**src**:	Lights, sounds, solar panels, motors — everything should be accessible.
	tgt:	Lichter, Töne, Solarelemente, Motoren — alles sollte verfügbar sein.
Cascaded	**src**:	Lights sounds solar panels motors everything should be accessible.
	tgt:	Licht klingt Solarpaneele, Motoren; alles sollte zugänglich sein.
XSTNet	**tgt**:	Licht, Geräusche, Solarkollektoren, Motoren — alles sollte zugänglich sein.
ConST	**tgt**:	Licht, Geräusche, Solarpanele, Motoren, alles sollte zugänglich sein.
		CASE 2
Ref.	**src**:	Eight years ago when I was at the Media Lab, I started exploring this idea of how to put the power of engineers in the hands of artists and designers.
	tgt:	Vor acht Jahren war ich am Media Lab und ich begann diese Idee zu erforschen, wie man die Macht der Ingenieure in die Hand von Künstlern und Designern legen könnte.
Cascaded	**src**:	Eight years ago when I was at the Media Lab, I started exploring this idea of how to put the power of engineers in the hands of artists and designers.
	tgt:	Vor 8 Jahren, als ich im Media Lab war, begann ich, diese Idee zu erforschen, wie man die Macht der Ingenieure in die Hände von Künstlern und Designern legte.
XSTNet	**tgt**:	Vor acht Jahren, als ich im Media Lab war, begann ich zu erforschen, wie man die Kraft der Ingenieure in die Hände von Künstlern und Designern legt.
ConST	**tgt**:	Vor acht Jahren, als ich im Media Lab war, begann ich, diese Idee zu erforschen, wie man die Macht von Ingenieuren in die Hände von Künstlern und Designern legt.

Table 7: En-De test cases that generated from the cascaded model, XSTNet (both provided by Ye et al. (2021)) and our ConST model.