Hindawi
Security and Communication Networks
Volume 2023, Article ID 9851426, 1 page
https://doi.org/10.1155/2023/9851426

Retraction

Retracted: Leveraging Multimodal Out-of-Domain Information to Improve Low-Resource Speech Translation

Security and Communication Networks

Received 26 December 2023; Accepted 26 December 2023; Published 29 December 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] W. Zhu, H. Jin, W. Yeh et al., "Leveraging Multimodal Out-of-Domain Information to Improve Low-Resource Speech Translation," *Security and Communication Networks*, vol. 2021, Article ID 9915130, 14 pages, 2021.

Hindawi
Security and Communication Networks
Volume 2021, Article ID 9915130, 14 pages
https://doi.org/10.1155/2021/9915130

Research Article

Leveraging Multimodal Out-of-Domain Information to Improve Low-Resource Speech Translation

Wenbo Zhu,[1] **Hao Jin**,[1] **WeiChang Yeh**,[2] **Jianwen Chen**,[3] **Lufeng Luo**,[3] **Jinhai Wang**,[1] **and Aiyuan Li**[4]

[1]*School of Mechatronic Engineering and Automation, Foshan University, Foshan 525000, Guangdong, China*
[2]*Integration and Collaboration Laboratory, Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 300, Taiwan*
[3]*School of Electronic Information Engineering, Foshan University, Foshan 525000, Guangdong, China*
[4]*Academic Journal Editorial Office, Foshan University, Foshan 525000, Guangdong, China*

Correspondence should be addressed to Wenbo Zhu; zhuwenbo@fosu.edu.cn

Received 17 September 2021; Accepted 28 October 2021; Published 26 November 2021

Academic Editor: Jian Su

Speech translation (ST) is a bimodal conversion task from source speech to the target text. Generally, deep learning-based ST systems require sufficient training data to obtain a competitive result, even with a state-of-the-art model. However, the training data is usually unable to meet the completeness condition due to the small sample problems. Most low-resource ST tasks improve data integrity with a single model, but this optimization has a single dimension and limited effectiveness. In contrast, multimodality is introduced to leverage different dimensions of data features for multiperspective modeling. This approach mutually addresses the gaps in the different modalities to enhance the representation of the data and improve the utilization of the training samples. Therefore, it is a new challenge to leverage the enormous multimodal out-of-domain information to improve the low-resource tasks. This paper describes how to use multimodal out-of-domain information to improve low-resource models. First, we propose a low-resource ST framework to reconstruct large-scale label-free audio by combining self-supervised learning. At the same time, we introduce a machine translation (MT) pretraining model to complement text embedding and fine-tune decoding. In addition, we analyze the similarity at the decoder side. We reduce multimodal invalid pseudolabels by performing random depth pruning in the similarity layer to minimize error propagation and use additional CTC loss in the nonsimilarity layer to optimize the ensemble loss. Finally, we study the weighting ratio of the fusion technique in the multimodal decoder. Our experiment results show that the proposed method is promising for low-resource ST, with improvements of up to +3.6 BLEU points compared to baseline low-resource ST models.

1. Introduction

Language translation has become an essential skill today. In this field, many technologies have also been generated, such as automatic speech recognition (ASR), machine translation (MT), and speech translation (ST). We know that the existing methods have achieved good results due to a large number of available speech and text resources. And these methods reach research and life practical standards. However, most nonmainstream languages do not have a sufficient training corpus due to the difficulties of collecting vocabulary and audio. And in low-resource work, the conventional methods do not give satisfactory results. Therefore, it is a challenge to improve the performance of ST tasks in low resources.

Traditionally, speech translation tasks include cascade and end-to-end structures. The cascade structures are based on jointly trained ASR [1] and MT [2] models. The advantage of this method is that it leverages text and audio resources to the greatest possible extent [3, 4]. In addition, the cascade model gives appropriate initial parameters for fine-tuning the ST task in the next step [5–8]. The end-to-end structure [9] is a separate subtask containing an ST encoder and an ST decoder. This approach is convenient for

joint tuning, and there is no need to require source text for training. However, end-to-end ST often cannot give competitive results in small sample tasks. Also, cascade methods in a low-resource context have been shown to learn phoneme quality better than end-to-end ST [10, 11]. In other words, phoneme quality can directly determine translation effectiveness. As a result, traditional low-resource ST tasks often use a cascade structure. However, cascade structures are easily prone to error propagation to leading to incorrect translations [12]. Therefore, in this paper, in low-resource ST tasks, we attempt to improve the error propagation problem that exists in cascade structures.

Usually, low-resource ST tasks focus on feature enhancement and model optimization. On the one hand, techniques such as data augmentation [10, 13–15], multitask learning [16–18], and pretraining of ASR data [8, 19–21] are used to enhance the feature representation. On the other hand, knowledge refinement [22], self-training [23], or multilingual ST [24–28] have been used to address actual scarcity in speech translation. Moreover, labeled ASR data, MT, or ST data provide additional information for multitask learning, pretraining, data augmentation, and multilingual ST. These techniques allow the model to learn more semantic information from unimodal data. In addition, both ASR and MT are conversions between the same data modality. As a result, several unimodal enhancement methods can effectively improve performance in ASR and MT tasks [29, 30]. However, ST is a task that translates source speech into text in the target language. It transforms from audio modality to text modality. The conversion of data modality indicates a higher complexity of the ST task. And in low-resource ST tasks, unimodal optimization does not achieve satisfactory results [31, 32]. Therefore, how to learn more common embeddings from enormous unlabeled out-of-domain multimodal data? How to avoid error propagation and achieve multimodal optimization for low-resource tasks? It is one of the motivations for our research.

Therefore, in this study, we use a cascade structure as the basis. On the one hand, we use self-supervised learning of many unlabeled out-of-domain acoustic representations to reconstruct. An unlabeled out-of-domain text pretraining model is also introduced to fine-tune the decoder. On the other hand, we studied the nonsimilarity on the decoder side. Then, we optimize the ensemble loss by using additional CTC loss and random pruning in the nonsimilar layer. These methods effectively solve the problem of error propagation and joint optimization. It is shown that the structure proposed in this paper can effectively combine out-of-domain audio and text data with improving the performance of low-resource ST tasks.

Our contributions are as follows:

(1) We propose a low-resource ST framework combined with self-supervised learning. And we analyze the effect of self-supervised learning on speech translation.

(2) We utilize decoder fusion techniques to fine-tune the overall model by introducing an out-of-domain unlabeled text pretraining model at the MT decoding end.

(3) We evaluated the decoded similarity and used random depth pruning to reduce the number of invalid pseudolabels to mitigate the problem of error propagation.

(4) We analyzed the nonsimilarity of decoding and added additional CTC loss to optimize the ensemble loss in the nonsimilar layer. It will better solve the multimodal optimization problem.

Experiments show that our optimal model can effectively utilize a large amount of unlabeled bimodal data to improve the performance of low-resource speech translation.

2. Related Work

This section discusses existing self-supervised learning and ST tasks that use textual information in the out-of-domain.

Self-supervised learning [33–35] is a machine learning (ML) paradigm that involves unsupervised learning of structural patterns of data using contextual data. It is prevalent for problems with small amounts of labeled data (for supervised training) and large amounts of unlabeled data (for self-supervised training). They have been proved successful in image classification, text classification, and NLP. In recent years, self-supervised learning has also proved effective for ASR tasks. The wav2vec is a self-supervised learning model. And the latest wav2vec 2.0 [36], both of which use contextual representations from the transformer model [37], was proposed by Facebook to learn to predict masked discrete speech codes. They both fine-tune on limited audio data to obtain good speech recognition results. In this paper, we use the wav2vec2.0 self-supervised ASR model as a basis to further investigate how to incorporate self-supervised learning applied to low-resource ST tasks.

Multitask learning by extracting information from this paper in the external domain has been widely used for ST tasks to overcome limited data [38–43]. However, studies have shown that multitask learning of single modal data may not apply to ST tasks with bimodal transitions. Standley et al. [44] conducted an empirical study on computer vision tasks for MTL. They found that "similar" tasks in MTL do not necessarily train better together. In addition, sequence-level knowledge distillation has been successfully shown to be applied to ST tasks. In a recent study, SeqKD was shown to reduce the demand for training data. Knowledge distillation refines the knowledge from one model to another by putting one [45]. Two-way SeqKD was proposed by Hirofumi Inaguma et al. [46]. It focuses on MT models from ex-domain textual resources. It also successfully demonstrates the efficacy of SeqKD in low-resource ST. This paper further investigates how to effectively combine a large amount of untagged bimodal information from the out-of-domain to improve the low-resource ST task.

Inspired by the above work, we investigate the practicality of multimodal techniques for low-resource speech translation and describe them in the following sections.

3. Methods

In this section, we describe a low-resource ST model based on a cascade structure. As shown in Figure 1, its framework consists of two independent subtasks: Automatic Speech Recognition (ASR) and Machine Translation (MT). Source audio $X_s = [x_1, x_2, \ldots, x_s]$. Generate source text $Y_s = [y_1, y_2, \ldots, y_s]$ by ASR task. Source text generates target text by MT task $Y_t = [y_1, y_2, \ldots, y_t]$.

In this paper, we introduce self-supervised learning on the audio coding side. To enhance the audio embedding representation by combining large-scale untagged out-of-domain audio information on small sample audio. And we present a multilingual text pretraining model to enrich text embedding at the decoding side to enhance text embedding.

However, it is easier to exacerbate the error propagation problem of cascade structures by performing multimodal optimization in two independent tasks. Therefore, optimizing the ST model for multimodal low-resource cascades and solving the error propagation problem is the most challenging task in this article. In this paper, we analyze the layer similarity at the decoding end. And random depth pruning is performed in a similar layer to reduce the model parameters and solve the model multimodal error propagation problem. To improve the multimodal optimization problem, we add an auxiliary intermediate layer of CTC loss in the nonsimilar layer to jointly optimize the model. This low-resource ST model can effectively combine many unlabeled bimodal extra-domain information. It enhances the modeling capability of low-resource ST models.

3.1. Encoder with Self-Supervised Learning

3.1.1. Baseline Architecture. Our approach uses conformer encoding as the baseline structure of the double encoding. Conformer is a multilayer attention architecture including self-attention and residual connectivity [47]. Self-attention learns global information. Residual connectivity helps train deep neural networks, where x is the input to the ASR encoder. L is the number of layers in the encoder. The $l-1$ layer to calculate the given information x_{l-1} is x_l:

$$\begin{aligned} x_l^{\mathrm{MHA}} &= \text{Self Attention}\left(x_{l-1}\right) + x_{l-1}, \\ x_l &= \text{Feed Forward}\left(x_l^{MHA}\right) + x_l^{\mathrm{MHA}}. \end{aligned} \tag{1}$$

The final representation is x_L, then fed to the standard CTC loss layer to optimize the audio alignment loss:

$$L_{\mathrm{CTC}} = -\log P_{\mathrm{CTC}}\left(y|x_L\right). \tag{2}$$

We also use SpecAugment technology. It enhances performance by strengthening the alignment of audio and text sequences in the form of speech spectrograms. To adjust the modeling scale, we set the kernel size of various CNNs to fit the acoustic representation of the model.

3.1.2. Self-Supervised Learning. Self-supervised learning uses auxiliary tasks to construct supervised information from large-scale unsupervised data automatically and train the network with such pseudolabels to learn representations that are valuable for downstream tasks. Therefore, this paper knows much unlabeled audio information in the out-of-domain to enrich the acoustic presentation by combining self-supervised learning. This paper combines the wav2vec2.0 self-supervised learning at the audio encoder side in the proposed low-resource ST system. This model reconstructs the acoustic representation of the out-of-domain information to improve the low-resource ST modeling capability.

The wav2vec2.0 model consists of a multilayer convolutional feature encoder f. The encoder consists of several blocks containing a time-domain convolution followed by layer normalization and a GELU activation function. It takes the original audio X as input and outputs the potential speech representation $Z_1, \ldots, Z_T$, i.e., $X \longrightarrow Z$, and the feature encoder's output to the transformer architecture's contextual network [48]. The dependencies of the potential representations $C_1, \ldots, C_T$ of the whole sequence are captured by self-attention to construct models to capture the information of the entire sequence [49], i.e., $Z \longrightarrow C$, where the contextual network uses a convolutional layer similar to [50, 51] as a relative position embedding instead of a fixed position embedding that encodes complete position information [52–54], where we compute the cosine similarity $\mathrm{sim}\,(a,b) = (\alpha^T b)/(\|\alpha\|\|b\|)$ between contextual representations and quantified latent speech representations. A quantified candidate representation $\widetilde{q} \in Q_t$, k distractors, and a true quantified potential speech representation q_t are the outputs of the contextual network.

$$L_c = -\log \frac{\exp\left(\mathrm{sim}\left(c_t, q_t\right)/k\right)}{\sum_{\widetilde{q\sim Q_t}} \exp\left(\mathrm{sim}\left(c_t, \widetilde{q}\right)/k\right)}. \tag{3}$$

Meanwhile, the wav2vec2.0 discretizes the output of the feature encoder, using the quantization module $Z \longrightarrow Q$ to represent the target in self-supervised training. For self-supervised training, the quantized representations are selected from multiple codebooks and linked together by quantization.

Given G a codebook, there are V entries $e \in R^{V \times d/G}$. We select one entry from each codebook, connect the obtained vectors, and apply a linear transformation $\mathrm{q} \in R^f$. Also, we use the straight-through estimator [55] and set G as the hard Gumbel softmax operation [56, 57]. The feature encoder output z is mapped to $\mathrm{l} \in R^{G \times V}$ logits, and the probability that the g group selects the V codebook entry is

$$p_{g,v} = \frac{\exp\left(\left(l_{g,v} + n_v\right)/\tau\right)}{\sum_{k=1}^{V} \exp\left(\left(l_{g,v} + n_k\right)/\tau\right)}, \tag{4}$$

where Γ is the nonnegative temperature, $n = -\log(-\log(u))$, and U is a uniform sample of (0,1). In the forward pass, the codewords i are $i = \mathrm{argmax}_i p_{g,j}$ selected, and in the backward pass, the true gradient of the Gumbel softmax output is used. In a batch of the corpus, the V entries in the G codebook are used on average by maximizing the entropy of the average softmax distribution1 of the G codebook entries for each codebook $\overline{p}_g$.

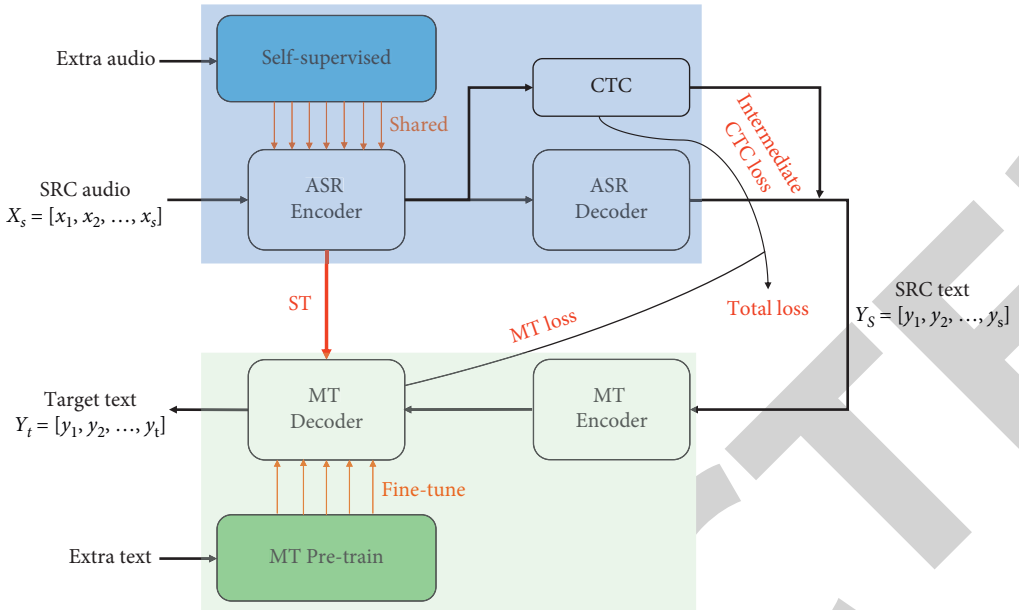

Figure 1: Structure of the integrated model.

$$L_d = \frac{1}{GV}\sum_{g=1}^{G} -H\left(\bar{p}_g\right) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}\bar{p}_{g,v}\log\bar{p}_{g,v}. \tag{5}$$

It learns the representation of speech audio during pretraining by solving the contrast loss L_c and the codebook diversity loss L_d where α is a tuned hyperparameter.

$$L = L_c + \alpha L_d. \tag{6}$$

We used raw 16-bit 16-kHz mono audio as the audio input in our experiments. We used the basic configuration of wav2vec2.0 to perform fine-tuning training on LibriSpeech's audio data, which contains fine-tuning models at different scales of 10 minutes, 100 hours, and 960 hours.

3.2. Decoder with Out-of-Domain Text Pretraining Model. To utilize large-scale unlabeled text data, this paper introduces an out-of-domain MT pretraining model to effectively use a large amount of unlabeled text data by fine-tuning it using a small amount of target domain text data. This paper achieves joint optimization by introducing a joint loss function to the dual model.

$$L(\theta) = -E_{x,y} \in D_{\mathrm{MT-Pretext}}\log P(y|x;\theta), \tag{7}$$

where θ is the model parameter and D is the target language text.

For the independent text generation work, we utilized the typical transformer-based approach. The decoder module has six transformer layers, of which layer 2048 is the most covert unit. We use prelayer normalization to make the training comparable, as the front-end model receives both speech representation and external text information as input.

Our experiments used the Adam optimizer with a learning rate of 2×10^{-4} and a warm-up of 25k steps. Based on the experimental results, MT pretraining provided a suitable introduction for the shared transformer module.

3.3. Multimodal Optimization Based on Hybrid CTC/Attention. The attention structure-based conformer decoder used in this paper is obtained entirely by data-driven, and the alignment relationship has no sequential restriction. In particular, in low-resource tasks, the lack of data volume will cause training difficulties for the attention structure-based model, and the alignment blindness will lead to a long training time. In contrast, the forward-backward algorithm of CTC can guide the output sequence to be aligned with the input sequence in the temporal order. Therefore, this paper adopts the hybrid CTC/attention model to avoid random alignment by CTC to speed up the training process.

3.3.1. Hybrid CTC/Attention. The training process of the hybrid CTC/attention model is multitask learning, combining CTC with the attention-based mechanism of cross-entropy $L(\mathrm{CTC})$ and $L(\mathrm{Att})$ and as shown in the following equation:

$$L = \lambda L(\mathrm{CTC}) + (1-\lambda)L(\mathrm{Att}), \tag{8}$$

where λ is a hyperparameter, usually less than 0.5 and usually taken as 0.2 or 0.3.

Among them, conjunction temporal classification (CTC) solves the sequence prediction problem by introducing monotonic alignment. For the encoder output $x \in R^{T\times D}$ where the length is T and the feature dimension is D , the CTC layer calculates the likelihood of the target sequence y:

$$P(y|x) \coloneqq \sum_{\alpha\in\beta^{-1}(y)} P(\alpha|x), \tag{9}$$

where $\beta^{-1}(y)$ is the set of possible alignments compatible with y and of length T, and α is the alignment in the set. The probability model of alignment is a factorization distribution:

$$P(\alpha[t]|x[t]) \coloneqq \prod_t P(\alpha[t]|x[t]),$$
$$P(\alpha[t]|x[t]) \coloneqq \text{Soft max}(\text{Linear}(x[t]))_{\alpha[t]}, \tag{10}$$

where $\alpha[t]$ and $x[t]$ are the t values of α and x, respectively. In the inference process, the most probable alignment is found by greedy decoding.

3.3.2. Multimodal Optimization. Performing multimodal optimization tasks may bring about problems such as model degradation and difficulty in cooptimization of multimodality [57]. Therefore, in this paper, we first analyze the similarity of the bimodal decoding end. And we introduce the random depth pruning technique in the similarity layer to mitigate the model degradation problem by selecting nonsimilar layers to assist the loss of additional intermediate layers of CTC. It will solve the problem that multimodal is challenging to optimize together.

(1) Random Pruning of Similar Layers. We analyze for the first time the decoding similarity of multimodal optimization in cascade tasks. And we use deep pruning techniques in the similarity layer to mitigate model degradation. Random depth [57–59] is a regularization method designed for deep residual networks [47]. After training the model with random depth, some layers are removed to obtain a new smaller submodel that does not require any fine-tuning and has good performance.

During training, each layer is skipped randomly with or without a given probability p. For each iteration, the $u = 1$ possibility of sampling from the Bernoulli distribution such that $u = 1$ is p and $u = 0$ is $1 - p$. Then, the remaining part is skipped (i.e., $x_l = x_{t-1}$). The output is calculated by modifying the equation at the decoding end (1) (2) as follows:

$$x_l^{\text{MHA}} = \frac{u}{p} \cdot \text{Self Attention}\left(x_{l-1}\right) + x_{l-1},$$
$$x_l = \frac{u}{p} \cdot \text{Feed Forward}\left(x_l^{\text{MHA}}\right) + x_l^{\text{MHA}}. \tag{11}$$

(2) Nonsimilar Layers for Auxiliary CTC Loss. We combine the analysis of nonsimilar and similar layers on a low-resource ST task for the first time and use an additional intermediate CTC loss in the nonsimilar layer. Intermediate CTC [58] is an auxiliary loss designed for CTC modeling. It regularizes the model using an additional CTC loss attached to the intermediate layer of the encoder. Let $l_1, \ldots, l_K$ be the intermediate layer and have a K position ($K < L$). The intermediate loss is defined as

$$L_{\text{InterCTC}} \coloneqq \frac{1}{K} \sum_k -\log P_{\text{CTC}}(y|x_{lk}). \tag{12}$$

The training objectives are then defined in conjunction with the above equation as

$$L \coloneqq (1 - w)L_{\text{CTC}} + wL_{\text{Inter CTC}}. \tag{13}$$

The intermediate representation is x_{lk}, with hyperparameters w. Note that the CTC and the intermediate CTC share the same linear projection layer of equation (11). The intermediate CTC is considered a regular CTC loss, just skipping all encoder layers after the intermediate layer. We use intermediate CTC loss in the nonsimilar layer.

In this paper, we first analyze the similarity at the decoding end. And we analyze the regularization achieved on the model by choosing single-stage and two-stage intermediate CTCs. We also discuss the impact of the choice of weighting parameters for both techniques on the experimental results.

4. Experimental Setting

This section describes our dataset for speech translation (ST), text data preprocessing, acoustic features, and optimizer setup. Then, we describe in detail how to train our baseline model.

4.1. Training Datasets. The dataset consists of Source and Target. The source dataset includes a speech dataset of Swahili variants, a small discourse dataset for both language pairings, and two parallel translated text corpora that make up the source dataset. The target dataset includes a single English speech translation dataset (see Table 1).

(A) Data for self-supervised models: LibriSpeech. To investigate the effect of unlabeled data on pretraining and self-training, we used 960 hours of LibriSpeech data for the wav2vec 2.0.

(B) Data for MT pretraining models: we selected two different MT models for different language translation tasks for the comparison experiments. For the MT pretraining model, we apply languid filtering, larger FFN, and resembling reverse translation finetuning, ensembling, and reordering to the text to obtain the optimal model.

4.2. Training and Decoding Details. Our implementation is based on the ESPnet-ST toolkit [59]. In the following, we provide details for reproducing the results. The pipeline is identical for all experiments.

(A) Baseline models: all experiments use the same 12-layer encoder structure, with a 6-layer decoder structure, an embedding size of 256, 4 attention heads, and FFN dimension of 2048. It includes 24 self-attentive blocks of dimension 1024, internal dimension 4096, and 16 attention heads for the self-supervised model. It generates a total of about 300 M parameters. In addition to the individual decoder model, we include an additional teacher network.

Table 1: Training datasets.

Language pair	SWA-EN	
	Train	Dev
Speech hours	5.3 h	1.9 h
Target tokens	985	—
Target utterances	4.5 K	868

The structure of the model is described in Section 3.3. We save the checkpoints with the best BLEU on the dev-set and average the last five checkpoints.

(B) Text preprocessing: transcriptions and translations were normalized and tokenized using the Moses tokenizer. All sources are marked in lowercase with punctuation removed, and targets are marked in the true case for fair comparison BLEU.

(C) Speech features: we used Kaldi [58] to extract 83-dimensional features (80-channel log Mel filterbank coefficients and 3-dimensional pitch features) that were normalized by the mean and standard deviation computed on the training set. For data augmentation, we used speed perturbation [60] with three factors of 0.9, 1.0, and 1.1 and SpecAugment [60] with three types of deterioration, including time warping (W), time masking (T), and frequency masking (F), where $W = 5$, $T = 40$, and $F = 30$.

(D) Optimization: we used the Adam optimizer [61] with the Noam [10, 62] learning rate schedule [63]. We set the initial learning rate to $1e-3$ and the dropout rate of 0.1. We used a batch size of 32 sentences per GPU, with a gradient accumulation of 2 and a clipping gradient of 5. As for model initialization, we trained two separate teacher models and used their weights to initialize the conformer model. We also included this shared model in the experiments. Finally, for decoding, we used a beam size of 10 with a length penalty of 0.6.

5. Results

In this section, we investigate the integration strategy of self-supervised learning and text pretraining on cascading ST. The performance and integration strategies of multimodal optimization techniques are also explored.

5.1. Baseline Work. The BLEU scores evaluated on the Swazi-English corpus are shown in Table 2. We use six groups of model baseline models as a comparison. Two of them are traditional models, including Seq2seq [63] and LSTM [64]. And they include four attention-based models, including attention-passing [65], transformer [55], transformer combined with knowledge distillation [66], and dual-encoding transformer [67]. In this paper, we use the conformer model as the basic structure, which is trained in cascade without using any textual resources of the source language.

Table 2: Comparison between baseline work and previous work.

Method	Train	Dev	BLEU Δ
Seq2seq	15.7	13.6	−2.4
LSTM	16.1	14.3	−0.7
Attention-passing	17.9	15.6	−0.4
Transformer	18.1	16.0	—
Transformer with SKD	18.4	15.9	−0.1
Dual-transformer	16.4	14.6	−1.4
Conformer kernel = 5 BPE 1k	17.2	14.6	−1.4
Conformer kernel = 7 BPE 1k	18.6	16.1	+0.1
Conformer kernel = 15 BPE 1k	19.4	17.3	+1.3

Inspired by previous work on segmenting speech into phone sequences based on phone change boundaries [68], we applied BPE-Dropout [69] with a rejection rate of 0.1 to reduce the phone sequence length. We believe that incorporating BPE-based phone features will give the model a deeper understanding of the sentences. Previous experiments have shown that the model's efficiency decreases when the BPE exceeds 1K, which may be related to the excessive granularity of the segmentation. Therefore, we use a BPE of 1K as the base segmentation element. And we choose the size of the convolution kernel to evaluate the best baseline working results.

5.2. Impact of Self-Supervised Data Scale on Encoding. In this section, we incorporate a self-supervised learning approach. The impact of different self-supervised on the low-resource cascade ST task is analyzed by examining the scale of audio information in the out-of-domain.

We observe significant gains using the wav2vec 2.0 model compared to the previous baseline (see Table 3). These baselines were not pretrained and did not use any additional other supervised speech translation data. The wav2vec 2.0 small model pretrained with 10 minutes of Librilight data achieved an average of 20.6 BLEU, which is 3.3 BLEU points better than the baseline average. These results show that the acoustic representation learned by the wav2vec 2.0 model is beneficial beyond speech recognition and applicable to speech translation. And it offers that the model proposed in this paper is combined with self-supervised learning. It can improve the ST task with insufficient source speech.

Compared to the previous baseline, we observed the attention weights of the conformer encoding combined with the self-supervised model. The combined self-supervised model was fine-tuned after 10 mins of data, without any other supervised speech translation data. The attentional alignment heat map for the encoded audio input and output is shown in Figures 2 and 3. The more diagonally correlated the weights are, the better the effect is on them (e.g., Figures 2 and 3), indicating the better learning ability of the encoder. Figure 3 shows that the attentional alignment ability of the self-supervised encoder attention weights is enhanced after fine-tuning compared to the baseline.

We observe (e.g., in Figure 4) the RTFs using different self-supervised models compared to the previous baseline. The baseline is not pretrained and does not use any other supervised

Table 3: Comparison of self-supervised models of different sizes with baseline models.

Model	Unlabeled Data	Labeled Scale	Train	Dev	BLEU Δ
Baseline	—	—	19.4	17.3	—
Wav2vec	LS-960	—	18.5	17.1	−0.2
Vq-wav2vec Gumbel	LS-960	—	18.9	17.2	−0.1
vq-wav2vec K-means	LS-960	—	18.7	16.9	−0.4
Wav2vec2.0 base	LS-960	No-fine-tune	19.1	17.6	+0.3
	LS-960	10 mins	25.9	19.3	+2.0
	LS-960	100 h	25.2	19.1	+1.8
	LS-960	960 h	24.5	19.1	+1.8
Wav2vec2.0 large	LS-960	No-fine-tune	19.2	17.6	+0.3
	LS-960	10 mins	26.7	20.6	+3.3
	LS-960	100 h	21.5	18.5	+1.2
	LS-960	960 h	17.8	17.4	+0.1

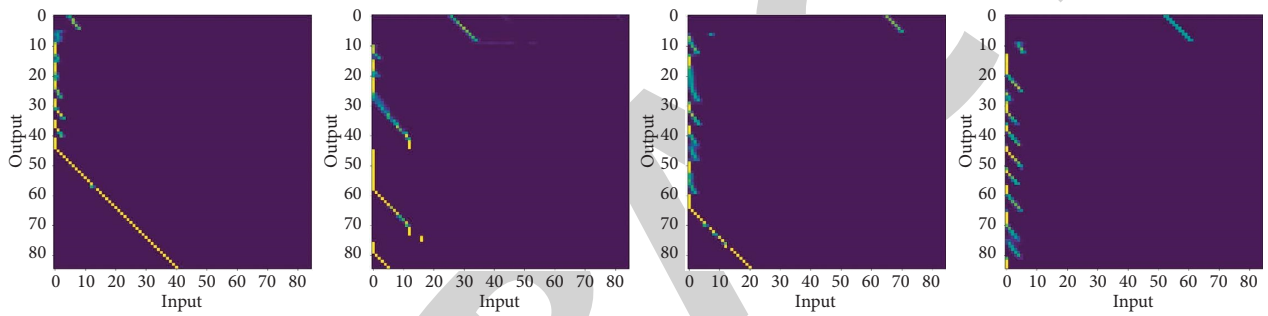

Figure 2: The alignment effect represents the baseline model encoder for the four input frames and the input text learned through the multiheaded attention mechanism. The better the alignment of the inputs and outputs is, the closer the lines of the heat map are to the diagonal.

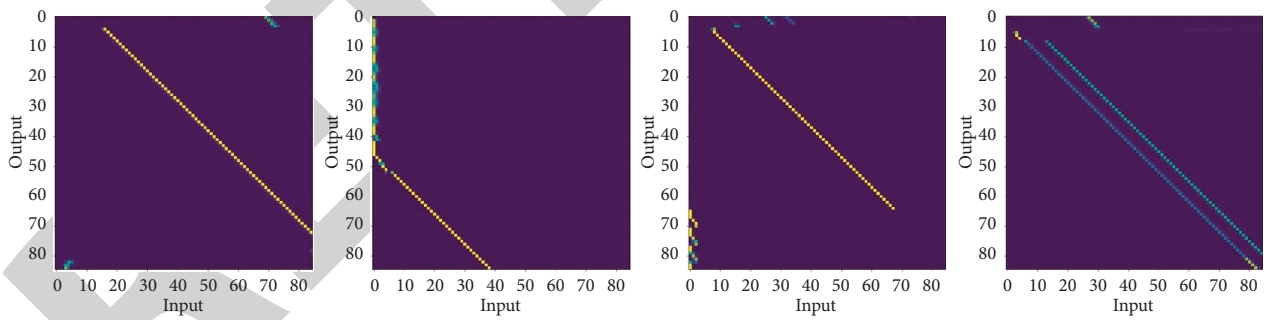

Figure 3: The alignment effect of the baseline model encoder combined with the self-supervised model for the four input frames and the input text through the multiple attention mechanism is indicated. Compared to Figure 2, the alignment effect of all four input frames and input text is improved.

speech translation data. The two different self-supervised models were fine-tuned using data of different sizes. Figure 4 shows that the RTF decreases for the two different self-supervised models as the size of the unlabeled data increases, and both are lower than the baseline model RTF.

5.3. Improvements from Decoding. Self-supervised learning uses unlabeled speech data to improve model performance. However, self-supervision generates noisy outputs that lead to models learning incorrect patterns. To inject more a priori knowledge of the target language, a good solution is to use the target domain's label-free text teacher model and fine-tune the student model on these. In this work, the pretrained model is improved by using additional unlabeled text from the language and using it to improve the generated decoding.

We use two different pretrained models of external MTs, one retaining only the single decoder and the other with the overall model in its entirety. These will accomplish

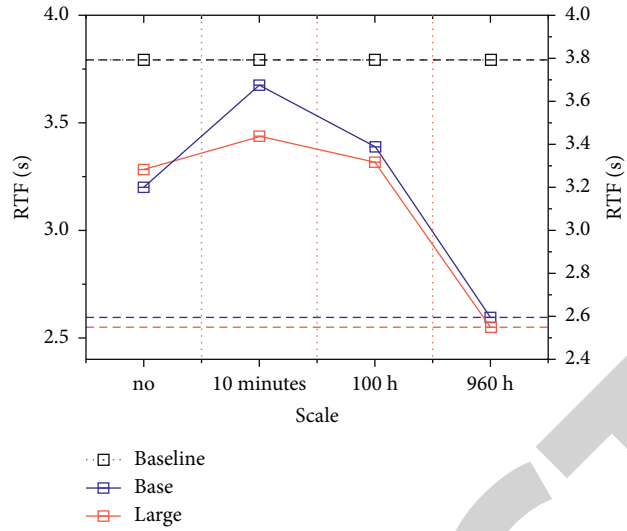

Figure 4: Impact of self-supervised models with the different labeled scales on RTF.

translation tasks of different sizes and in different languages. They are used to evaluate the impact of external MT on BLEU scores. The results show (see Table 4) that the pretrained models for MT tasks in different languages effectively improve the performance of the baseline models. The best performance of the single model is improved by 1.4 BLEU. The 2.0 BLEU improved the best performance of the dual model. The results show that additional text untagged information in the same target language versus different target languages can help the ST task decode rich text embeddings.

Compared to the previous baselines, we observed the attention weights using different MT pretrained models. These baselines were self-supervised fine-tuned with 10 minutes of labelable data. Figures 5 and 6 show the self-supervised model without incorporating out-of-domain text pretraining and the model with incorporating out-of-domain text pretraining, respectively. The experiments show (e.g., in Figures 5 and 6) that the text attention weights aligned well with the ex-domain text pretraining model. It implies that the ex-domain text MT pretraining model can effectively improve the performance of the low-resource ST task.

5.4. Improvements of Leveraging Multimodal Information

5.4.1. Impact of Similar Layer Pruning. We trained a 24-layer hybrid CTC/attention model, and we explored the effect of different depth layer pruning rates on the number of parameters and RTF. We set four different configurations: (a) depth-rate = 0.1, (b) depth-rate = 0.2, (c) depth-rate = 0.3, (d) depth-rate = 0.4 and (e) depth-rate = 0.5 compared with the baseline model, as in Figure 7.

We observe (e.g., in Figure 7) the RTF of the self-supervised model using different random depth pruning rates. Experiments show that the number of model parameters and the RTF is reduced using the random depth pruning technique compared to the previous baseline. To a certain extent, the model degradation problem is solved.

Based on the above four settings, we explored the effect of layer pruning on BLEU at different depths in Table 5.

Experiments show (see Table 5) that the performance is similar for random depths of 0.2 and 0.4. This can be explained by the similarity of layer four and layer eight at the decoding end and the higher learning ability of the relevant layers. However, the BLEU is the lowest at layer two and layer 10. This represents the most robust learning ability of the audio representation in the relevant layer. In summary, compared to the baseline model, the best results are obtained when we use a random depth of 0.3. The BLEU effect is minimal, the model parameters are reduced, and the RTF decreases. In summary, when the random depth pruning is taken as 0.3, it helps to reduce the overall number of parameters and solve the model degradation problem to some extent.

5.4.2. Impact of Auxiliary Losses in Nonsimilar Layers. First, we explored the positional variation of the intermediate CTC and concluded that it has a minimal impact on accuracy. Accordingly, the results show that random depth pruning helps to reduce the model parameters. However, we found that the model performance did not improve. Therefore, we further optimize the bimodal data by introducing the additional intermediate CTC loss by the similar layer selected above.

We use CTC-assisted loss with different layers to determine the effect of layers on common loss. We fine-tune the self-supervised model using 10 minutes of labeled data compared to the previous baseline. And we decode and fine-tune the text model using the out-of-domain MT pretraining.

The experiments show (e.g., in Figure 8) that for models with different layers of additional losses, the auxiliary CTC loss is at two layers. In particular, the best losses were achieved at 6 and 12 layers.

Table 4: Comparison of different decoder teacher models to the baseline model.

Encoder	Decoder	Decoder-pretrain	Fine-tune	Description	Train	Dev	BLEU Δ
Wav2vec2.0 large	Conformer	No	No	—	27.1	20.9	—
	Conformer	transformer.wmt19.single	Yes	En⟶De	26.8	21.4	+0.5
				De⟶En	27.0	22.1	+1.2
				En⟶Ru	26.7	21.2	+0.3
				Ru⟶En,	27.4	22.3	+1.4
	Conformer	transformer.wmt19	Yes	En⟶De ensemble	26.2	21.6	+0.7
				De⟶En ensemble	28.4	22.9	+2.0
				En⟶Ru ensemble	27.1	22.2	+1.3
				Ru⟶En ensemble	27.8	22.4	+1.5

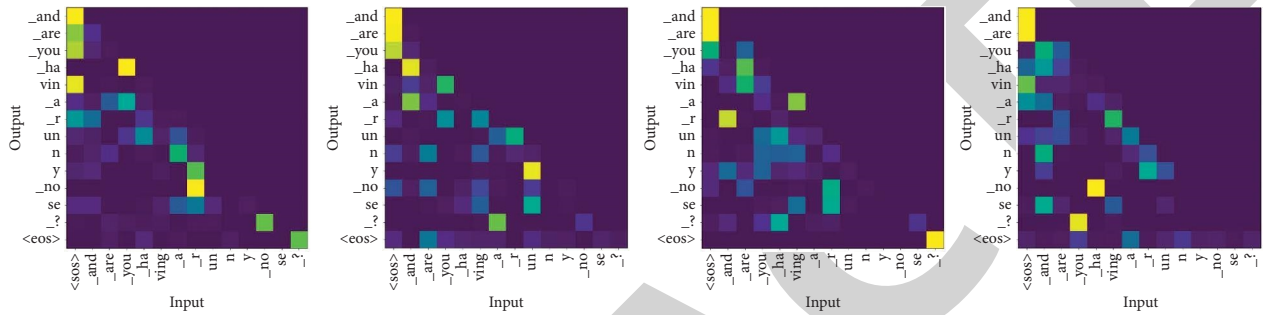

Figure 5: An alignment effect represents the four decoded output texts of the baseline model learned through the self-attention mechanism. The better the alignment effect of the input and output is, the closer the lines of the heat map are to the diagonal.

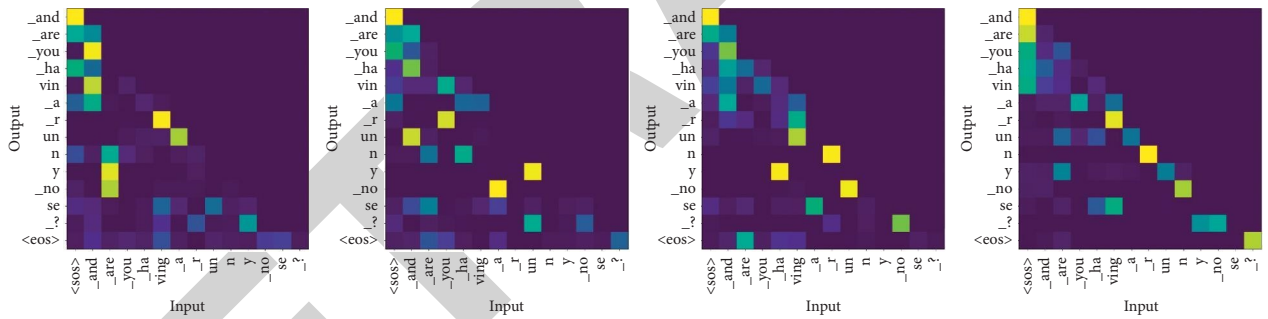

Figure 6: The alignment effect represents the four decoded output texts of the baseline model combined with the out-of-domain text pretraining model learned through the self-attention mechanism. The better the alignment effect of the input and output is, the closer the lines of the heat map are to the diagonal.

We use the effect of CTC-assisted loss without location on multimodal optimization. Compared to previous baselines, these were fine-tuned with 10 min of labeled data and decoded and fine-tuned using a dual-model out-of-domain MT pretrained text model.

Experiments show (see Table 6) that the best results are obtained when we use the intermediate CTC. The recognition effect improves, and the BLEU decreases. It can be explained by the learned representation being better for networks with layers 6 and 12, which do better with auxiliary loss. The performance is similar for the intermediate CTC layers 12 and 18. It can be explained by the fact that the similarity is consistent for all layers below layer 12. However, the worst effect is when at layer 18, representing that the auxiliary loss does not learn the commonality well. In summary, the intermediate CTC helps to improve the performance of the low-resource ST model for multimodal optimization [67–71].

5.4.3. Impact of Related Weight. For the above experiments, it is a challenging task to combine the intermediate CTC and the follow-on depth pruning effectively. The effect of both is further investigated by introducing correlation weights. First, the experiments show that the intermediate CTC has good performance at encoders 1/4 and 1/2. Therefore, we added an intermediate layer at layer 12 and layer 6, respectively.

We use five sets of weight parameters separately to compare the experimental effects. Table 7 shows that the increase of weights contributes to the overall similarity. The overall product is most satisfactory when $w = 0.66 \approx 2/3$. When the weight parameters continue to increase, it brings poor results. It indicates that the learning ability of the layer is limited at this point. Finally, Table 7 shows that combining the two regularizations greatly increases the similarity of the layers. This indicates that the effective combination of the

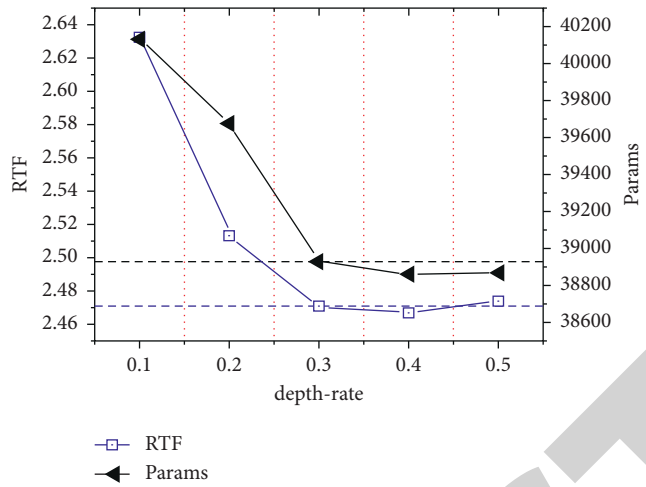

Figure 7: The effect of self-supervised models with different random depth pruning rates on RTF and parameters.

Table 5: Comparison of the self-supervised and baseline models with different random depth pruning rates.

Model	Unlabeled data	Labeled scale (mins)	Depth-rate	Train	Dev	BLEU Δ
Wav2vec2.0 large	LS-960	10	—	26.7	20.6	—
			0.1	25.9	19.9	−0.7
			0.2	26.0	20.1	−0.5
			0.3	26.3	20.4	−0.2
			0.4	26.1	20.3	−0.3
			0.5	25.8	20.0	−0.6

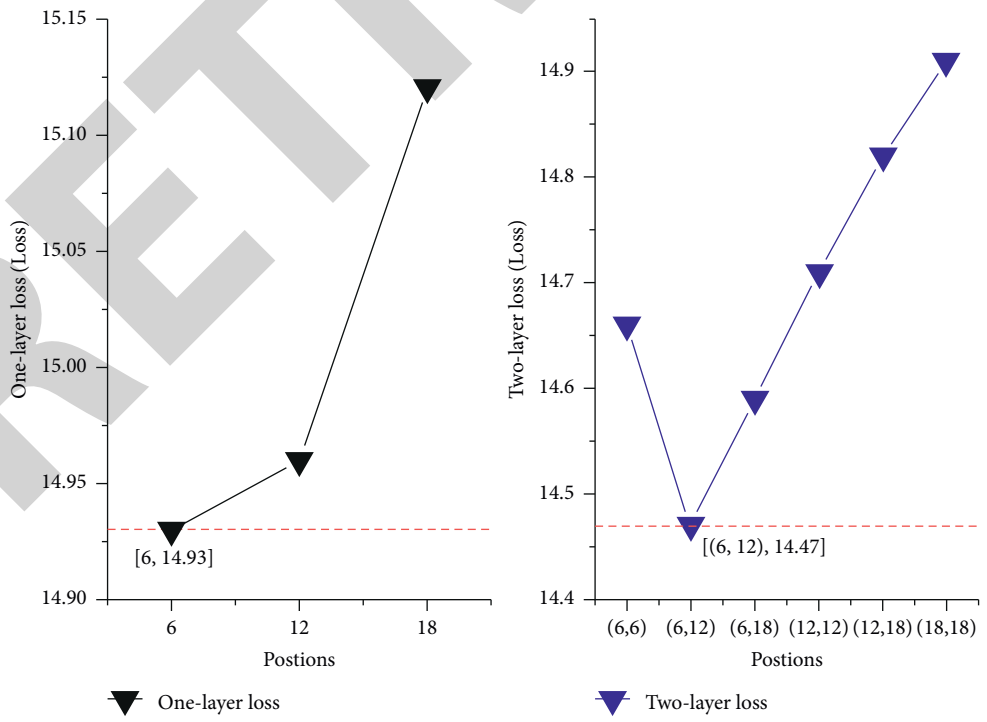

Figure 8: Effect of self-supervised models with different locations on the common loss.

Table 6: Comparison of the CTC-assisted loss at different positions with the baseline model.

Model	Unlabeled data	Labeled scale (mins)	Position	Train	Dev	BLEU Δ
Wav2vec2.0 large	LS-960	10	—	26.7	20.6	—
			(6)	26.1	20.4	−0.2
			(12)	26.2	20.0	−0.6
			(18)	25.9	20.1	−0.5
			(6, 6)	26.4	20.4	−0.2
			(6, 12)	26.6	20.5	−0.1
			(6, 18)	26.3	20.4	−0.2
			(12, 12)	26.1	20.1	−0.5
			(12, 18)	26.4	20.3	−0.3
			(18, 18)	26.0	20.2	−0.4

Table 7: Comparison of CTC-assisted loss and random depth pruning combining different weights of the self-supervised model with the baseline model is shown.

Model	Unlabeled data	Labeled scale	Depth-rate	Postion	Weight	Train	Dev	BLEU Δ
Wav2vec2.0 large	LS-960	10 mins	0.3	(6, 12)	0	26.3	20.4	—
		—	—	—	0.25	26.5	20.3	−0.1
		—	—	—	0.33	26.7	20.4	0
		—	—	—	0.5	26.9	20.6	+0.2
		—	—	—	0.66	27.1	20.9	+0.5
		—	—	—	1	26.6	20.5	+0.1

two regularization methods will fully utilize the likeness of the intermediate layers with each layer. And it helps to reduce the number of parameters and optimize the overall low-resource ST model.

6. Discussion

Although this study demonstrates the effectiveness of bimodal learning for low-resource ST, it does not explore the deep-level relationship between speech and text. And how can we further combine multimodality? These are enormous challenges.

7. Conclusion

In the low-resource ST challenge, we learn by combining self-supervised and text pretraining methods. On average, the result of the earlier approach is improved by 2 BLEU in the low-resource ST task. We also analyze the similarity at the decoding end. And we use random depth pruning in the similarity layer to mitigate the degradation of the model. Also, an additional CTC-assisted loss is used in the non-similar layer to optimize the merging loss. It further improves the BLEU by 0.5. Our study proposes an innovative approach for speech translation with low resources.

Data Availability

We used the IBAN dataset, which is a publicly accessed dataset (https://www.openslr.org/24).

Conflicts of Interest

All authors declare no conflicts of interest.

Authors' Contributions

Wenbo Zhu and Hao Jin conceptualized the study; Hao Jin developed the methodology; WeiChang Yeh was responsible for the software; Wenbo Zhu, Jianwen Chen, and Jinhai Wang validated the study; Hao Jin did formal analysis; Wenbo Zhu investigated the study; WeiChang Yeh provided resources; Jianwen Chen curated the data; Hao Jin wrote the original draft; Wenbo Zhu reviewed and edited the manuscript; Lufeng Luo visualized the study; WeiChang Yeh supervised the study; Aiyuan Li did project administration; Wenbo Zhu was responsible for funding acquisition. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This thesis would never have materialized without the help and support. First, the authors would like to express their sincere gratitude to their supervisor at China Foshan University, Professor Zhu Wenbo, who gave them a lot of valuable and constructive advice on their thesis. With his professional and academic knowledge, he taught them how to do research and revise the theory. Whenever the authors sent him an e-mail concerning their view, he replied soon. He spent a lot of time reading and correcting their idea. Only under his guidance and encouragement could the authors finish this thesis. The authors also indebted to all their teachers in the laboratory. As teachers , the authors learned a lot about interpretation, translation, culture, and teaching methods, which is helpful to their job. At last, the authors want to thank their family members and relatives, who showed their concern and support when the authors pursued their study. This work was supported by the National

Natural Science Foundation of China, Youth Program (No. 62106048), the Basic and Applied Basic Research Project of Guangdong Provincial Science and Technology Department-Guangdong-Foshan Joint Fund Program (No. 2019A1515110273), the Beijing Jiaotong University Key Laboratory of Rail Transit Control and Safety, Open Program (No. RCS2019K010), and the Natural Science Foundation of Guangdong Provincial Science and Technology Department (No. 2017KTSCX194).

References

[1] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pp. 4960–4964, IEEE, Shanghai, China, March 2016.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, Cornell University, Ithaca, New York, Mar 2015.

[3] A. Anastasopoulos and D. Chiang, "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*, vol. 1, pp. 82–91, ACL Anthology, New Orleans, Louisiana, June 2018.

[4] Y. Liu, J. Zhang, H. Xiong et al., "Synchronous speech recognition and speech-to-text translation with interactive decoding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 8417–8424, PKP Publishing Services Network, April 2020.

[5] Y. Liu, H. Xiong, Z. He et al., "End-to-end speech translation with knowledge distillation," *Proc. Interspeech*, pp. 1128–1132, 2019.

[6] S.-P. Chuang, T.-W. Sung, A. H. Liu, and H.-y. Lee, "Worse WER, but better BLEU? Leveraging word embedding as intermediate in multitask end-to-end speech translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5998–6003, Association for Computational Linguistics, July 2020.

[7] C. Wang, Y. Wu, S. Liu, M. Zhou, and Z. Yang, "Curriculum pre-training for end-to-end speech translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3728–3738, Association for Computational Linguistics, July 2020.

[8] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9161–9168, PKP Publishing Services Network, April 2020.

[9] A. B'erard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: a proof of concept for end-to-end speech-to-text translation," in *Proceedings of the NIPS workshop on End-to-end Learning for Speech and Audio Processing*, Cornell University, Ithaca, New York, Dec 2016.

[10] E. Salesky, M. Sperber, and A. W. Black, "Exploring phoneme level speech representations for end-to-end speech translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, July 2019.

[11] E. Salesky and A. W. Black, "Phone features improve speech translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2388–2397, Association for Computational Linguistics, July 2020.

[12] N. Ruiz and M. Federico, "Assessing the impact of speech recognition errors on machine translation quality," in *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 261–274, Association for Machine Translation in the Americas, Vancouver, Canada, October 2014.

[13] Y. Jia, M. Johnson, W. Macherey et al., "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, UK, May 2019.

[14] J. Pino, L. Puzon, J. Gu, X. Ma, A. D. McCarthy, and D. Gopinath, "Harnessing indirect training data for end-to-end automatic speech translation: tricks of the trade," in *Proceedings of the IWSLT*, Cornell University, Ithaca, New York, October 2019.

[15] A. D. McCarthy, L. Puzon, and J. Pino, "Skinaugment: auto-encoding speaker conversions for automatic speech translation," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, IEEE, Barcelona, Spain, May 2020.

[16] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proceedings of the Interspeech 2017*, pp. 2625–2629, Cornell University, Ithaca, New York, October 2017.

[17] A. B' erard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Calgary, AB, Canada, April 2018.

[18] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A general multitask learning framework to leverage text data for speech to text tasks," in *Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021.

[19] S. Bansal, H. Kamper, K. Livescu, L. Adam, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of the North*, pp. 58–68, Association for Computational Linguistics, Minneapolis, MN, USA, June 2019.

[20] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of the North*, Association for Computational Linguistics, Minneapolis, MN, USA, June 2019.

[21] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pre-training for low-resource speech-to-text translation," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[22] Y. Liu, H. Xiong, Z. He et al., "End-to-end speech translation with knowledge distillation," in *Proceedings of the Interspeech*, pp. 1128–1132, ISCA, Berlin, Germany, April 2019.

[23] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-training for end-to-end speech translation," in *Proceedings of the Interspeech*, pp. 1476–1480, ISCA, Berlin, Germany, October 2020.

[24] M. A. Di Gangi, M. Negri, and M. Turchi, "One-to-many multilingual end-to-end speech translation," in *Proceedings of*

the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 2019.

[25] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilin-gual end-to-end speech translation," in *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 2019.

[26] C. Wang, J. Pino, A. Wu, and J. Gu, "Covost: a diverse multilingual speech-to-text translation corpus," in *Proceedings of the Interspeech, 2021*, pp. 2247–2251, ICSA, Berlin, Germany, April 2020.

[27] C. Wang, A. Wu, and J. Pino, *Covost 2 and Massively Multilingual Speech-To-Text Translation*, arXiv, Ithaca, NY, USA, 2020.

[28] X. Li, C. Wang, Y. Tang et al., *Multilingual Speech Translation with Efficient fine-tuning of Pre-trained Models*, arXiv, Ithaca, NY, USA, 2020.

[29] Q. Xu, A. Baevski, T. Likhomanenko et al., "Self-training and pre-training are complementary for speech recognition," in *Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021.

[30] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Association for Computational Linguistics, Minneapolis, Minnesota, July 2020.

[31] C. Wang, Y. Wu, Y. Qian et al., "Unispeech: unified speech representation learning with labeled and unlabeled data," in *Proceedings of the of ICML*, Cornell University, Ithaca, New York, Jun 2021.

[32] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pre-training for low-resource speech-to-text translation," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[33] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[34] A. v. d. Oord, Y. Li, and O. Vinyals, *Representation Learning with Contrastive Predictive Coding*, arXiv, Ithaca, NY, USA, 2018.

[35] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Sori-cut, "Albert: A lite BERT for self-supervised learning of language representations," in *Proceedings of the International Conference on Learning Representations*, Cornell University, Ithaca, NY, USA, Sep 2019.

[36] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., , 2020.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al., *Attention Is All You Need*, arXiv, Ithaca, NY, USA, 2017.

[38] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Association for Computational Linguistics, Austin, Texas, November 2016.

[39] R. J. Weiss, C. Jan, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proceedings of the INTERSPEECH*, pp. 2625–2629, ISCA, Berlin Germany, 2017.

[40] A. Anastasopoulos and D. Chiang, "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*, vol. 1, pp. 82–91, Association for Computational Linguistics, New Orleans, Louisiana, June 2018.

[41] P. Bahar, B. Tobias, and N. Hermann, "A comparative study on end-to-end speech to text translation," in *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 2019.

[42] S. Reddy Indurthi, H.J Han, N. Kumar Lakumarapu et al., "End-end speech-to-text translation with modality agnostic meta-learning," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[43] C. Wang, Y. Wu, S. Liu, Y. Zhenglu, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9161–9168, PKP Publishing Services Network, Burnaby, Canada, April 2020.

[44] X. Li, C. Wang, Y. Tang et al., *Multilingual Speech Translation with Efficient fine-tuning of Pre-trained Models*, arXiv, Ithaca, NY, USA, 2020.

[45] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multitask learning?" in *Proceedings of the International Conference on Machine Learning*, Cornell University, Ithaca, New York, Jul 2020.

[46] H. Inaguma, T. Kawahara, and S. Watanabe, "Source and target bidirectional knowledge distillation for end-to-end speech translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1872–1881, Association for Computational Linguistics, New Orleans, Louisiana, June 2018.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[48] Y. Liu, M. Ott, N. Goyal et al., *A Robustly Optimized Bert Pre-training Approach*, arXiv, Ithaca, NY, USA, 2019.

[49] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020.

[50] A. Mohamed, D. Okhonko, and L. Zettlemoyer, *Transformers with Convolutional Context for ASR*, arXiv, Ithaca, NY, USA, 2019.

[51] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, "Payless attention with lightweight and dynamic convolutions," in *Proceedings of the of ICLR*, Cornell University, Ithaca, New York, Feb 2019.

[52] F. Wen, J. Shi, and Z. Zhang, "Closed-form estimation algorithm for EMVS-MIMO radar with arbitrary sensor geometry [J]," *Signal Processing*, vol. 186, p. 108117, 2021.

[53] G. Zheng, Y. Song, and C. Chen, "Height measurement with meter wave polarimetric MIMO radar: signal model and MUSIC-like," *Algorithm [J]*, vol. 190, p. 108344, 2022.

[54] X. Wang, L. T. Yang, D. Meng, D. Mianxiong, O. Kaoru, and W. Huafei, "Multi-UAV cooperative localization for marine targets based on weighted subspace fitting in SAGIN environment," *IEEE Internet of Things Journal*, 2021.

[55] E. Jang, S. Gu, and B. Poole, *Categorical Reparameterization with Gumbel-Softmax*, arXiv, Ithaca, NY, USA, 2016.
[56] D. Jiang, X. Lei, W. Li et al., *Improving Transformer-Based Speech Recognition Using Unsupervised Pre-training*, arXiv, Ithaca, NY, USA, 2019.
[57] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," in *Proceedings of the Interspeech*, May2019.
[58] J. Lee and S. Watanabe, "Intermediate loss regularization for ctc-based speech recognition," in *Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021.
[59] A. Fan, E. Grave, and A. Joulin, "Reducing transformer depth on demand with structured dropout," in *Proceedings of the ICLR*, Cornell University, Ithaca, New York, May 2020.
[60] S. Watanabe, T. Hori, S. Karita et al., "ESPnet: end-to-end speech processing toolkit," in *Proceedings of the Interspeech, 2018*, pp. 2207–2211, ISCA, Berlin Germany, Mar 2018.
[61] D. Povey, A. Ghoshal, G. Boulianne et al., "The Kaldi speech recognition toolkit," in *Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, Montreal, Canada, Jan2011.
[62] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, pp. 3586–3589.
[63] D. S. Park, W. Chan, Y. u Zhang et al., "Specaugment: a simple data augmentation method for automatic speech recognition," in *Proceedings of the Interspeech 2019*, pp. 2613–2617, ISCA, Graz, Austria, February2019.
[64] D. P. Kingma and B. Jimmy, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, Cornell University, San Diego, CA, USA, May 2015.
[65] R. J. Weiss, C. Jan, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to sequence models can directly translate foreign speech," in *Proceedings of the Interspeech*, pp. 2625–2629, ISCA, Graz, Austria, February2019.
[66] A. Berard, L. Besacier, A. Can Ko-cabiyikoglu, and P. Olivier, "End-to-end automatic speech translation of audiobooks," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6224–6228, IEEE, Calgary, AB, Canada, April 2018.
[67] M. Sperber, N. Graham, N. Jan, and A. Waibel, "Attention-passing models for robust and data-efficient end-to-end speech translation," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 313–325, 2019.
[68] Y. Liu, H. Xiong, Z. He et al., "End-to-end speech translation with knowledge distillation," in *Proceedings of the InterSpeech*, pp. 1128–1132, ISCA, Graz, Austria, February2019.
[69] L. Hang, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Dual decoder transformer for joint automatic speech recognition and multilingual speech translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3520–3533, International Committee on Computational Linguistics, Barcelona, Spain, March 2019.
[70] A. B. Rico Sennrich and H. Barry, "Neural machine translation of rare words with sub-word units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Association for Computational Linguistics, Berlin, Germany, August 2016.
[71] I. Provilkov, D. Emelianenko, and E. Voita, "BPE-dropout: simple and effective subword regularization," in *Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1882–1892, Association for Computational Linguistics, Berlin, Germany, July 2020.