EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY Open Access

Adaptive multi-task learning for speech to text translation

Xin Feng[1], Yue Zhao[1*], Wei Zong[1] and Xiaona Xu[1]

Abstract

End-to-end speech to text translation aims to directly translate speech from one language into text in another, posing a challenging cross-modal task particularly in scenarios of limited data. Multi-task learning serves as an effective strategy for knowledge sharing between speech translation and machine translation, which allows models to leverage extensive machine translation data to learn the mapping between source and target languages, thereby improving the performance of speech translation. However, in multi-task learning, finding a set of weights that balances various tasks is challenging and computationally expensive. We proposed an adaptive multi-task learning method to dynamically adjust multi-task weights based on the proportional losses incurred during training, enabling adaptive balance in multi-task learning for speech to text translation. Moreover, inherent representation disparities across different modalities impede speech translation models from harnessing textual data effectively. To bridge the gap across different modalities, we proposed to apply optimal transport in the input of end-to-end model to find the alignment between speech and text sequences and learn the shared representations between them. Experimental results show that our method effectively improved the performance on the Tibetan-Chinese, English-German, and English-French speech translation datasets.

Keywords Speech to text translation, Optimal transport, Multi-task learning, Cross attentive regularization

1 Introduction

A speech-to-text translation (ST) system is commonly a pipeline framework, which consists of two components, an automatic speech recognition (ASR) model and a machine translation (MT) model [1, 2]. The source language speech is transcribed by the speech recognition model, and then the transcribed text is translated by the MT model into target language text. However, such cascaded models suffer from error propagation and high latency. Recent works proposed an end-to-end speech translation (E2E ST) model [3, 4], which provides an effective solution by jointly optimizing a single model for conversions from source language speech to target language text. Although the E2E ST model has the advantages above, its special nature as a cross-modal and cross-language task introduces a challenge–data scarcity. Therefore, present research usually leverages knowledge acquired from MT tasks to assist in the training of ST models.

For low-resource languages, multi-task learning frameworks are commonly used to achieve knowledge sharing between tasks, thereby improving the performance of the model on the target task. The performance of a multi-task model on each task tends to improve as the corresponding weight assigned to it increases. However, when the weights exceed a certain threshold, the model's performance gradually decreases [5]. Different combinations of weights lead to variations in model performance. Therefore, it is an important concern how to allocate weights for each task in order to achieve optimal performance on the target task. There are two typical methods for

*Correspondence:
Yue Zhao
zhaoyueso@muc.edu.cn
[1] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Zhongguancun South Street, Beijing 100081, China

adjusting multi-task weights. One approach is to manually assign weights to each task and continuously experiment with various weight combinations in search of the best combination. The other approach is to use dynamic adjustment techniques [6]. Compared to manual adjustment of task weights, dynamically adjusting them during training allows for faster and more efficient identification of optimal weight combinations [7, 8]. In this paper, we use the adaptive cross-entropy loss function based on task loss proportion as our multi-task objective function. This method is not only feasible but also achieves better allocation for weights.

Additionally, due to the modality gap between speech and text, ST cannot learn MT knowledge well, resulting in ST performance often lagging behind MT tasks. Previous studies have proved that when the input speech representation is similar to its corresponding text representation, information is better transferred from MT task to ST task, leading to improved ST performance [9]. Therefore, we proposed obtaining representations of speech and text that are close to each other in Wasserstein [10] space by optimal transport (OT) methods to reduce the gap between the speech representation and the corresponding transcription. By the cross-language conversion ability learned in MT tasks to ST tasks, the ST model can learn the better correspondence between source language speech and target language text with a small amount of parallel corpus data.

Our contributions are as follows: (1) a weight-updating scheme based on loss proportion is adopted to dynamically adjust the weights of each task during model training. Thus, the adaptive ability of the multi-task ST model is improved. (2) Based on the multi-task training framework, we introduce the cross-modal optimal transport method for ST, which reduces the gap between speech representation and corresponding transcription. (3) Experimental results on public speech translation datasets show that the proposed method can significantly improve the model performance.

2 Related work

2.1 End-to-end ST

To overcome error propagation and reduce latency of cascade ST systems, Bérard et al. [3] proposed to use an end-to-end architecture to directly translate speech into text in another language without intermediate transcription, which has become the dominant paradigm in recent years. However, the development of ST has been hampered by the scarcity of ST data and the cross-modal and cross-language characteristics. To address this problem, researchers usually use pre-training [11–13], multi-task learning [14–16], and knowledge distillation [17, 18] to introduce additional data and other tasks to improve performance.

2.2 Multi-task learning

Multi-task learning aims to enhance the target task by using related auxiliary tasks. Although multi-task learning is effective, manually adjusting the weights of each task is indeed a tedious task. Therefore, dynamic adjustment of weights is usually used, which can be broadly categorized into two types: gradient-based methods and loss-based methods. Among the gradient-based methods, Chen et al. [7] studied the gradients from different tasks and conduct task dependent gradient normalization to encourage different tasks to learn at similar speed. For loss-based methods, Kendall et al. [5] weighed multiple loss functions by taking into account the mean square error uncertainty of each task. Liu et al. [8] proposed the dynamic weighted average (DWA) method, which uses the average of task losses over time to measure task losses. However, these methods usually add extra complexity to the training phase. In this paper, we employ a weight updating scheme based on loss proportions for automatically adjusting multi-task weights.

2.3 Optimal transport

OT is a classical mathematical problem. It is commonly used to describe the transfer cost between two distributions. Villani et al. [19] provided a systematic and comprehensive exposition of the OT theory. In recent years, this theory has been widely used in research to find consistency between languages or modalities. Chen et al. [20] used OT in image-text pre-training to achieve fine-grained alignment between words and image regions. Gu et al. [21] used OT to bridge the gap between semantically equivalent representations of different languages in the field of MT, and Zhou et al. [22] used OT to integrate two modal representations that are mixed to overcome the modality gap between speech and text to improve the performance of ST. Compared to this approach, in this paper, we obtain representations of speech and text close to each other in the Wasserstein space through OT to reduce the gap between the speech representation and the corresponding transcription.

2.4 Bridging the modality gap

It is still difficult to fully use MT data using the above techniques due to the modal differences between speech and text. Several works have attempted to bridge this gap. Liu et al. [23] reduced the length of speech representations to match text representations and narrowed the representation gap by minimizing their L2 distance; Xu et al. [13] mapped speech representations to text representations by connecting temporal classification and mapping layers; Fang et al. [24] blended sequences of speech and text representations in order to bridge the modality gap; Han et al. [25] projected speech and text features into a shared semantic space; Zhou et al. [22] mixed speech and text sequences across modalities through optimal transport; and Ye et al. [9] brought sentence-level representations closer together through contrast learning. Different from previous studies, in this paper, we reduce the modality gap from the embedding representation between speech and text, design effective methods to learn similar representations of speech and text, and establish connections between different perceptual modalities, so that the ST model can better use information from different modalities, and ultimately improve the performance.

3 Methods

In this section, we will first describe the method of reducing the modality gap between speech and text through optimal transport (OTST). And then adaptive multi-task learning for OTST is introduced in detail. Based on E2E Transformer, we leverage the Wasserstein distance between the speech feature sequence and the text feature sequence using the optimal transport before the Transformer encoder and add the OT loss to the model training loss to make the encoded speech and its corresponding text close to each other in the Wasserstein space. In model training, three weights assigned to ST, MT, and ASR loss are automatically tuned according the proportion. Figure 1 provides a schematic depiction delineating the conceptual framework of our proposed methodology.

3.1 Problem formulation

Speech translation aims to translate source language speech into target language text. The corpus of ST is usually composed of triplet data $D = \{(s, x, y)\}$, where $s = (s_1, \ldots, s_{|s|})$ represents the source language speech sequence, $x = (x_1, \ldots, x_{|x|})$ is the transcript from the source language, and $y = (y_1, \ldots, y_{|y|})$ is the corresponding translation in the target language, $|s|$, $|x|$ and $|y|$ respectively represent their lengths.

3.2 Model architecture

We use the same multi-task network model architecture as XSTNet [26], which combines multiple training tasks of ST, ASR and MT, aiming to achieve E2E ST. The model consists four modules: a speech encoder, a text embedding layer, a Transformer [27] encoder, and a Transformer decoder. It supports audio and text inputs, and these two inputs share the Transformer module in the model.

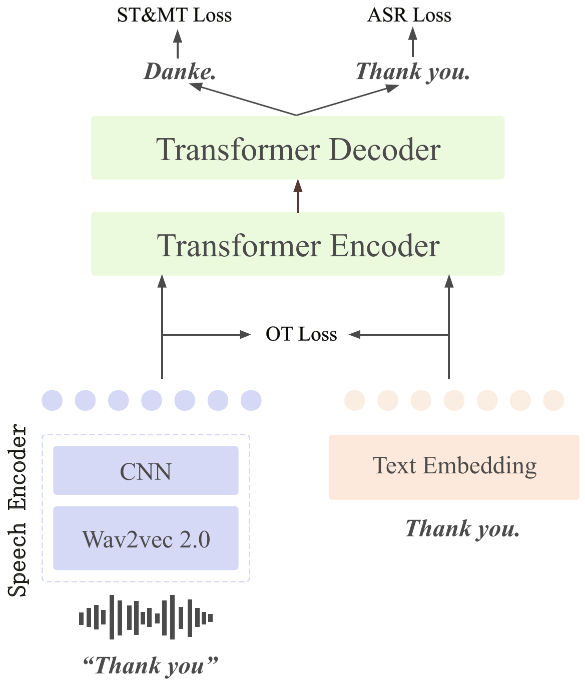

Fig. 1 The model structure of adaptive OTST

Speech encoder extracts contextualized acoustic embeddings from the raw waveform. It consists of Wav2vec 2.0 [28] and subsampler. The input is raw waveform signal sampled at 16 kHz. Wav2vec 2.0 first extracts a speech representation from the original waveform signal, but the output sequence of Wav2vec 2.0 is usually much longer than the corresponding text sequence. To further match the length of the audio representation and text sequence, we further add 2 convolutional layers with stride of 2 after Wav2vec 2.0 to reduce the time dimension of the speech representation by a factor of 4.

Text embedding is set in parallel with the speech encoder to capture semantic information in the text and map the text token into embeddings. We calculate the Wasserstein distance from the parallel speech and text sequences obtained from the speech encoder and text embedding layer through optimal transport.

Moreover, both the speech encoder and the text embedding layer are connected to the Transformer encoder. The encoder receives the output of the speech encoder or text embedding layer and further learns semantic information, which is then processed by the Transformer decoder to obtain the final output of the model.

We first undertake pre-training of the model using external MT data and then optimize the entire model by minimizing cross entropy loss.

$$\mathcal{L}_{ST} = -\sum_{n} \log P(y_n|s_n) \tag{1}$$

Given the model can support speech and text inputs, we introduce auxiliary MT and ASR tasks during the training process to obtain multi-task cross-entropy loss $\mathcal{L}_{multi-ce}$. Finally, we formulate the training objective as $\mathcal{L} = \mathcal{L}_{multi-ce} + \lambda\mathcal{L}_{ot}$, where$\mathcal{L}_{ot}$ represents the Wasserstein distance between speech and text sequences, and λ denotes the weight parameter governing $\mathcal{L}_{ot}$.

3.3 Cross-modal optimal transport

Optimal transport (OT) is a classical mathematical problem that provides powerful tools for comparing different probability distributions [29]. It is usually the solution to the problem of minimizing the cost of transferring one distribution to another, so that the distance between two discrete probability distributions is minimized after transmission. If we regard speech and text sequences as two independent distributions, OT can be used to measure the distance between them.

3.3.1 Optimal transport

For two discrete probability distributions α and β,

$$\alpha = \sum_{i=1}^{n} a_i \delta_{u_i} \qquad s.t. \sum_{i=1}^{n} a_i = 1 \tag{2}$$

$$\beta = \sum_{j=1}^{m} b_j \delta_{v_j} \qquad s.t. \sum_{j=1}^{m} b_j = 1 \tag{3}$$

where, α is represented by the mass $a_1, \cdots, a_i, \cdots, a_n \in [0, \infty)$ at position $u_1, \cdots, u_i, \cdots, u_n \in \mathbb{R}^d$, δ_{u_i} is the value of the Dirac delta function at position u_i, β is represented by the mass $b_1, \cdots, b_j, \cdots, b_m \in [0, \infty)$ at position $v_1, \cdots, v_j, \cdots, v_m \in \mathbb{R}^d$, δ_{v_i} is the value of the Dirac delta function at position v_i.

Given the transportation cost function $c(U_i, V_j)$, let $Z_{ij} \geq 0$ represent the mass transferred from U_i to V_j, then the total transportation cost can be expressed as $\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} C(U_i, V_j)$. Let Z^* be the transportation plan with the lowest transportation cost, which is calculated as follows:

$$\min_{Z} \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} c(u_i, v_j) = \min_{Z} < C, Z > \quad s.t. \, Z \geq 0, \sum_{j=1}^{m} Z_{ij} = a_i, \sum_{i=1}^{n} Z_{ij} = b_j \tag{4}$$

where Z and C denote the $n \times m$ matrices whose elements are Z_{ij} and $C_{ij} = c(u_i, v_j)$.

3.3.2 Wasserstein distance

The Wasserstein distance between α and β is defined as $W(\alpha, \beta) =< C, Z^* >$ but evaluating it is expensive in practice. Usually, the upper-bound approximation function $W_\lambda(\alpha, \beta)$ for the Wasserstein distance is solved, defined as

$$W_\lambda(\alpha, \beta) = \min_{Z} < C, Z > -\lambda H(Z) \tag{5}$$

where $H(Z) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(Z_{ij}) \log(p(Z_{ij}))$ is the entropy function, which is used as a regularization to improve the optimization result. $\lambda > 0$ is a regularization weight. $p(Z_{ij})$ denotes the probability of passing Z_{ij} units of mass from position u_i to position v_j. W_λ is evaluated using the Sinkhorn algorithm [30].

3.3.3 Wasserstein distance between speech and text

For the two independent distributions of speech and text sequences, we can use OT to measure the distance between them. Set the speech sequence as $H^s = (h_1^s, \ldots, h_i^s, \ldots, h_n^s)$ and the text sequence as $H^x = \left(h_1^x, \ldots, h_j^x, \ldots, h_m^x\right)$. Define two distributions α and β, whose mass is uniformly distributed at positions $(h_1^s, \ldots, h_i^s, \ldots, h_n^s) \in \mathbb{R}^d$ and $\left(h_1^x, \ldots, h_j^x, \ldots, h_m^x\right) \in \mathbb{R}^d$, that is, the mass at all positions of distribution α is $\frac{1}{n}$, and the mass at all positions of distribution β is $\frac{1}{m}$. Let the transportation cost of a unit mass from h_i^s to h_j^x be $C\left(h_i^s, h_j^x\right) = \|h_i^s - h_j^x\|_p$, with $p \geq 1$ (typically p=2). $\mathcal{L}_{ot} = W_\lambda(\alpha, \beta)$ can be seen as the difference between speech and text sequences, and we call this value the Wasserstein distance, which is added as a loss to the model training loss function.

3.4 Adaptive cross-entropy loss

Multi-task cross-entropy loss $\mathcal{L}_{multi-ce} = \omega_1\mathcal{L}_{ST} + \omega_2\mathcal{L}_{ASR} + \omega_3\mathcal{L}_{MT}$, where $\mathcal{L}_{ST}$, $\mathcal{L}_{ASR}$, and $\mathcal{L}_{MT}$ are cross-entropy losses on $< s, y >$, $< s, x >$, and $< x, y >$ pairs. The weight ω_1, ω_2, and ω_3 correspond to the extent to which the model updates each task during the training process. The cross-entropy loss functions for ST task, ASR task, and MT task are as follows:

$$\mathcal{L}_{ST} = -\sum_{n} \log P\left(y_n|s_n\right) \tag{6}$$

$$\mathcal{L}_{ASR} = -\sum_{n} \log P(x_n|s_n) \tag{7}$$

$$\mathcal{L}_{MT} = -\sum_{n} \log P(y_n|x_n) \tag{8}$$

To allocate task weights more effectively and optimize the model's performance on the target task, the weight ω at training step t is determined by the proportion of the corresponding loss value at training step $t-1$ to the total loss value. We express the weight as:

$$\omega_1(t) = \frac{\mathcal{L}_{ST}(t-1)}{\mathcal{L}_{ST}(t-1) + \mathcal{L}_{ASR}(t-1) + \mathcal{L}_{MT}(t-1)} \tag{9}$$

$$\omega_2(t) = \frac{\mathcal{L}_{ASR}(t-1)}{\mathcal{L}_{ST}(t-1) + \mathcal{L}_{ASR}(t-1) + \mathcal{L}_{MT}(t-1)} \tag{10}$$

$$\omega_3(t) = \frac{\mathcal{L}_{MT}(t-1)}{\mathcal{L}_{ST}(t-1) + \mathcal{L}_{ASR}(t-1) + \mathcal{L}_{MT}(t-1)} \tag{11}$$

Therefore, the model can dynamically adapt its learning strategy according to the learning level of each task, thereby find the optimal combination of task weight to balance multi-task learning. Finally, the weight update scheme based on the loss proportion is referenced into the multi-task cross-entropy loss function, and the loss for training steps t is obtained as:

$$\mathcal{L}_{multi-ce}(t) = \omega_1(t)\mathcal{L}_{ST}(t) + \omega_2(t)\mathcal{L}_{ASR}(t) + \omega_3(t)\mathcal{L}_{MT}(t) \tag{12}$$

4 Experiments

4.1 Datasets

4.1.1 ST datasets

We evaluate our methods presented in this paper on Tibetan-Chinese (Ti-Zh), English-German (En-De), and English-French (En-Fr) directions. The Ti-Zh dataset was constructed from the TIBMD@MUC [31] dataset. For the En-De and En-Fr directions, we used the MuST-C [32] dataset from TED Talks. The detailed statistics of the dataset are shown in Table 1.

4.1.2 External MT datasets

We also introduce external MT datasets to pre-train our translation model. For En-De and En-Fr directions, we randomly selected 2 million and 250,000 bilingual parallel sentences from the WMT [33] dataset, respectively. For Ti-Zh directions, We collated 270,000 bilingual parallel sentences based on the TIBMD@MUC dataset.

4.2 Experimental setups

4.2.1 Model configuration

Our implementation is based on the FAIRSEQ toolkit [34]. Following the standard practices in ST, we employ the Wav2vec 2.0 model with overlaid subsamplers as the speech encoder. The subsampler consists of two convolutional layers with a stride of 2, kernel size of 5, and an output channel size of 512, aimed at reducing the length of the speech sequence and alleviating the length discrepancy between speech and text embeddings. The dimensionality of the text embedding layer is set to 512. For the Transformer, we adopt a basic configuration, including 6 layers for both the encoder and decoder, each layer comprising 512 hidden units, 8 attention heads, and 2048 feed-forward network (FFN) hidden states.

4.2.2 Data preprocessing

For the speech input, we use the 16-bit 16 kHz mono-channel raw audio. To ensure training efficiency, we filter out samples with frames greater than 480k or less than 1k. As for the text input, we tokenize transcripts and translations using the SentencePiece [35] model. The vocabulary size of 10k is shared between source and target languages. For external MT datasets, parallel sentence pairs with length ratios exceeding 1.5 are filtered out.

4.2.3 Experimental details

During the training phase, we employ the Adam optimizer [36] to update parameters, with an initial learning rate set to 2×10^{-4} and warm-up steps to 15k, dropout of 0.1. In the inference phase, we use beam search with a beam size of 10. We evaluate the BLEU on the test set using sacreBLEU [37] as the evaluation metric for the translation task. All models are trained on Nvidia V100 GPUs.

Table 1 Statistics of the dataset

Datasets	Ti-Zh		En-De		En-Fr	
	Hours	Sents	Hours	Sents	Hours	Sents
train	64.04	27340	94	56000	52.68	30000
valid	1.92	1000	2	1423	2.55	1412
tst	1.86	1000	4	2641	4	2403

Table 2 BLEU scores of different models

Models	Ti-Zh(w/o)	Ti-Zh(w/)	En-De(w/o)	En-De(w/)	En-Fr(w/o)	En-Fr(w/)
base	12.71	12.84	18.87	22.78	22.60	27.42
XSTNet	12.79	13.00	20.61	23.00	24.38	28.48
ConST	13.40	13.63	20.77	23.10	24.39	27.92
STEMM	13.36	13.61	20.82	23.13	24.36	27.89
OTST	13.47	13.90	20.88	23.09	25.19	28.59
adaptive-OTST	**13.88**	**14.00**	**20.91**	**23.19**	**25.42**	**28.61**

Table 3 Results of each model under extremely low-resource settings

Models	BLEU		
	Ti-Zh	En-De	En-Fr
base	5.59	6.29	12.47
XSTNet	5.11	6.51	14.40
ConST	5.68	6.52	15.66
OTST	5.98	6.65	15.98
adaptive-OTST	**6.06**	**7.30**	**16.11**

5 Results

Table 2 shows the BLEU values of each model. Compared to the base model, adaptive-OTST has an improvement of 1.16, 0.41, and 1.19 BLEU in the Ti-Zh, En-De, and En-Fr directions, respectively. We also compare our approach with other baseline models, including XSTNet [26] using progressive training procedure, ConST [9] using contrastive learning strategy, and STEMM [24] using mixed speech representation sequences and word embedding sequences. As most existing performance improvements rely on the utilization of large-scale external MT data, for fair comparison, we study two settings: (1) without external MT data and (2) with external MT data. For settings without external MT data, our method improved 2.01 BLEU on average in three directions compared to base model. For settings with external MT data adaptive-OTST's performance also surpasses that of other strong baselines.

To validate our method under extremely low-resource settings, we constructed 10 hours ST subsets using random sampling from the TIBMD@MUC Tibetan-Chinese dataset and the MuST-C dataset, respectively. In the extremely low-resource ST setting, we compared our method with other models, with results shown in Table 3. Adaptive-OTST consistently outperforms baseline methods in all three language directions.

Table 4 Ablation study in Ti-Zh direction

Exp.	Config.	BLEU
I	adaptive-OTST	**14.00**
II	$-\mathcal{L}_{ASR}-\mathcal{L}_{MT}$	11.63
III	$-\mathcal{L}_{ot}$	13.30
IV	– dynamic weight adjustment	13.90

6 Analysis

6.1 Ablation study

As a multi-task learning framework, the performance of our ST system is influenced by the training objectives. Through ablation study, this paper evaluates the impact of multi-task learning modules and OT methods on model performance. Table 4 shows the performance of the model under different training objectives. The experimental results indicate that for the Ti-Zh translation direction, both the multi-task learning module and the OT method contribute to the improvement of model performance. Based on the results of Exp I, Exp II, and Exp IV, it can be concluded that using multi-task learning methods can bring an improvement in 2.37 BLEU, removing adaptive cross-entropy loss in multi-task dynamic weight adjustment methods results in a decrease in translation performance. The results of Exp I and Exp III indicate that the introduction of OT loss methods can achieve significant performance advantages in ST systems.

6.2 Comparison between OT and other losses

In this paper, we reduce the distance between speech and text representations by introducing the OT method. In order to prove the effectiveness of the OT method, we introduce cross-attentive regularization (CAR) [38] at the input layer of the Transformer encoder.

Due to the distinct input modalities of speech and text, their representations may have different lengths and cannot be directly compared. Hence, we first reconstruct the speech feature sequence from the output of the speech

Table 5 BLEU scores for different losses

Extra loss	BLEU
OT loss	**14.00**
CAR loss	13.20

encoder and the text feature sequence from the output of the text embedding layer. The two reconstructed sequences are calculated from the text output sequence via self-attention or the speech output sequence via cross attention over the text output sequence. Both reconstructed sequences have the same length and the similarity between the speech and text feature sequences can be measured by the L2 distance between these two reconstructed sequences, where a smaller distance indicates higher similarity between speech and text.

Table 5 shows the BLEU scores of the models under different methods, with the OT loss resulting in a 0.88 BLEU higher score than the CAR loss.

6.3 Positions of OT

In speech translation, speech features can be simply divided into acoustic features and semantic features. After the speech signal is processed by a speech encoder, the number of acoustic and semantic features in the low-level speech representation is roughly equivalent. However, in the high-level speech representation output by the Transformer encoder, semantic features usually dominate. The modal differences between these two layers of speech representation and their corresponding text representation are apparent. However, there is currently no consensus on which layer's modal difference reduction would yield a more significant impact on enhancing the performance of ST models. To this end, we introduced OT techniques in the input and output layers of the Transformer encoder to reduce modal differences. As shown in Table 6, introducing OT in the input layer of the Transformer encoder to reduce modal differences can result in better performance of the model compared to the output layer. We believe that there is more original alignment information in the lower layers of the model, which is more suitable for OT calculation.

6.4 Weight setting for OT loss

In this section, we discuss the impact of the OT loss weight λ. We experimented with several λ values ranging from 0.1 to 1.0. Figure 2 visually demonstrates the variation in model BLEU scores with different λ values, with the highest BLEU achieved when $\lambda = 0.25$. When the OT loss weight λ is too small, its effectiveness in reducing modality gap is minimal, resulting in the model's inability to effectively leverage MT for improving ST performance. Conversely, when λ is too large, the model excessively focuses on narrowing the modality gap between speech and text, leading to a decline in the performance of the primary task ST. Therefore, we opted for a moderate weight setting, selecting the hyperparameter $\lambda = 0.25$ to achieve optimal model performance.

Table 6 BLEU scores with different positions of OT

OT	BLEU
Encoder in	**14.00**
Encoder out	13.12

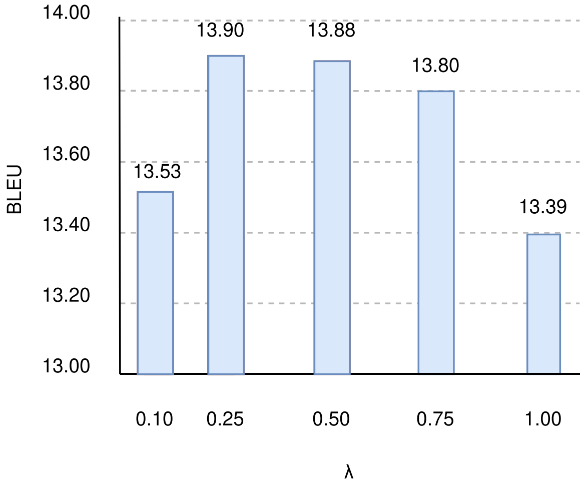

Fig. 2 BLEU scores under different weights λ

7 Conclusion

In this paper, we propose adaptive-OTST, which uses an adaptive cross-entropy loss function based on task loss proportion as the multi-task objective function to improve the adaptive ability of the multi-task ST model. In addition, it reduces the modality gap by bringing closer the distance between speech and text representations in the Wasserstein space, leading to better performance. The experiment demonstrates the efficacy of our approach in low-resource ST. In the future, we hope to integrate the optimal transport with other methods to bridge the modality gap and further improve the performance of ST.

Acknowledgements
Not applicable.

Authors' contributions
XF proposed algorithm ideas, conducted experiments, and was a major contributor in writing the manuscript. YZ offered guidance, formulated the research plan, and refined the paper. WZ and XX provided valuable insights and suggestions for improving the paper's content. All authors read and approved the final manuscript.

Funding
This work was supported by funds as follows: (1) the National Natural Science Foundation of China (NSFC), "Research on end-to-end multi-task learning-based multi-dialect speech recognition method for Tibetan" (61976236), (2) enterprise-funded project "Research and realization of end-to-end sequence data modeling."

Availability of data and materials
The datasets to this study are accessible in designated repositories. The must-c datasets utilized in this research are obtainable from the [FBK] repository at [https://mt.fbk.eu/must-c-releases/]. The Tibetan-Chinese datasets utilized herein were compiled from the [OpenSLR] repository, [http://www.openslr.org/124/]. Additionally, the external machine translation datasets employed in this study can be accessed through [https://www.statmt.org/wmt16/translation-task.html].

Declarations

Competing interests
The authors declare no competing interests.

Received: 18 April 2024 Accepted: 25 June 2024
Published online: 13 July 2024

References

1. F.W.M. Stentiford, M.G. Steer, Machine translation of speech. Br. Telecom Technol. J. **6**(2), 116–122 (1988)
2. A. Waibel, A.N. Jain, *ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, JANUS: A speech-to-speech translation system using connectionist and symbolic processing strategies (Toronto, 1991), pp. 793–796
3. A. Bérard, O. Pietquin, C. Servan, Listen and translate: A proof of concept for end-to-end speech-to-text translation. CoRR. (2016). arXiv:1612.01744
4. L. Duong, A. Anastasopoulos, *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, An Attentional Model for Speech Translation without Transcription (Association for Computational Linguistics (ACL), Stroudsburg, 2016), pp. 949–959
5. A. Kendall, Y. Gal, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics (IEEE Computer Society, Washington, DC, 2018), pp. 7482–7491
6. W. Vandenhende, S. Georgoulis, Multi-task learning for dense prediction tasks: A survey. IEEE Trans. Pattern. Anal. Mach. Intel. **44**(7), 3614–3633 (2021)
7. Z. Chen, V. Badrinarayanan, in *Proceedings of the International Conference on Machine Learning*, Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks (Proceedings of Machine Learning Research (PMLR), Cambridge, 2018), pp. 14–16
8. S. Liu, E. Johns, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, End-to-end multi-task learning with attention (IEEE Computer Society, Washington, DC, 2019), pp. 1871–1880
9. R. Ye, M. Wang, L. LI, Cross-modal contrastive learning for speech translation. Phys. Lett. 5099–5113 (2022). arXiv:2205.02444
10. C. Frogner, C. Zhang, Learning with a Wasserstein loss. Adv. Neural Inf. Process. Syst. **28**, 2053–2061 (2015)
11. A. Alinejad, A. Sarkar, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, Effectively pretraining a speech translation decoder with machine translation data (Association for Computational Linguistics (ACL), Stroudsburg, 2020), pp. 8014–8020
12. R. Zheng, J. Chen, *International Conference on Machine Learning*, Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation (PMLR, 2021). pp. 12736–12746
13. C. Xu, B. Hu, Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders (2021), pp. 2619–2630. https://doi.org/10.18653/v1/2021.acl-long.204
14. H. Le, J. Pino, C. Wang, Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation (2020), pp. 3520–3533. arXiv:2011.00747
15. H.K. Vydana, M. Karafiát, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jointly trained transformers models for spoken language translation (IEEE Signal Processing Society, Piscataway, 2021), pp. 7513–7517
16. Y. Tang, J. Pino, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, A general multi-task learning framework to leverage text data for speech to text tasks (IEEE Signal Processing Society, Piscataway, 2021), pp. 6209–6213
17. M. Gaido, M.A. Di Gangi, M. Negri, End-to-end speech translation with knowledge distillation: FBK@ IWSLT2020 (2020), pp. 80–88. arXiv:2006.02965
18. H. Inaguma, T. Kawahara, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Source and Target Bidirectional Knowledge Distillation for End-to-end Speech Translation (Association for Computational Linguistics (ACL), Stroudsburg, 2020), pp. 1872–1881
19. C. Villani, *Optimal transport: Old and new* (Springer, Berlin, 2009)
20. Y.C. Chen, L. Li, L. Yu, *European conference on computer vision*, Uniter: Universal image-text representation learning (Springer International Publishing, Cham, 2020), pp. 104–120
21. S. Gu, Y. Feng, in *findings of the Association for Computational Linguistics: EMNLP*, Improving zero-shot multilingual translation with universal representations and cross-mappings (Association for Computational Linguistics (ACL), Stroudsburg, 2022), pp. 6492–6504
22. Y. Zhou, Q. Fang, in *proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, CMOT: Cross-modal mixup via optimal transport for speech translation (Association for Computational Linguistics (ACL), Stroudsburg, 2023), pp. 7873–7887
23. Y. Liu, J. Zhu, Bridging the modality gap for speech-to-text translation (2020). arXiv:2010.14920
24. Q. Fang, R. Ye, in *proceedings of the 60th Annual Meeting of the Association for Computational Lingui stics*, Stemm: Self-learning with speech-text manifold mixup for speech translation (Association for Computational Linguistics (ACL), Stroudsburg, 2022), pp. 7050–7062
25. C. Han, M. Wang, H. Ji, Learning shared semantic space for speech-to-text translation. CoRR. 2214–2225 (2021). arXiv:2105.03095
26. R. Ye, M. Wang, End-to-end speech translation via cross-modal progressive training 2267–2271 (2021). https://doi.org/10.21437/INTERSPEECH.2021-1065
27. A. Vaswani, N. Shazeer, Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 5998–6008 (2017)
28. A. Baevski, Y. Zhou, Wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
29. G. Peyré, M. Cuturi, Computational optimal transport: With applications to data science. Found. Trends® Mach. Learn. **11**(5–6), 355–607 (2019)
30. M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport. Adv. Neural Inf. Process. Syst. **26**, 2292–2300 (2013)
31. Y. Zhao, X. XU, An open speech resource for Tibetan multi-dialect and multitask recognition (OpenSLR, 2020). http://www.openslr.org/124/. Accessed 22 June 2023
32. M.A. Di Gangi, R. Cattoni, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Must-c: A Multilingual Speech Translation Corpus (ELSEVIER SCI LTD, Oxon, 2019), pp. 2012–2017
33. O. Bojar, R. Chatterjee, *First conference on machine translation*, Findings of the 2016 Conference on Machine Translation (wmt16) (Association for Computational Linguistics (ACL), Stroudsburg, 2016), pp. 131–198
34. C. Wang, Y. Tang, Fairseq S2T: Fast speech-to-text modeling with Fairseq (2020), pp. 33–39. arXiv:2010.05171
35. T. Kudo, J. Richardson, in *Proceedings of the 2018 Conference on Empirical Methods in Nat ural Language Processing: System Demonstrations*, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing (2018), pp. 66–71 2018. eprint arXiv:1808.06226,cs.CL
36. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization (2014). arXiv:1412.6980

37. M. Post, *Proceedings of the Third Conference on Machine Translation*, A call for clarity in reporting BLEU scores (2018), pp. 186–191. eprint arXiv:1804.08771, cs.CL
38. Y. Tang, J. Pino, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Improving speech translation by understanding and learning from the auxiliary text translation task (Association for Computational Linguistics (ACL), Stroudsburg, 2021), pp. 4252–4261

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.