Tibetan-Chinese Speech-to-Speech Translation Based on Large Speech Models

Hetong Fan[1,2], Yue Zhao[1,2†], Jing Yu[1,2], Zairan Gong[1,2], Xiaona Xu[1,2], Haizhou Li[3]

[1]*Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China*

[2]*Minzu University of China, Beijing 100081, China*

[3]*The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China*

Abstract

There are currently three main approaches to implementing a speech-to-speech translation system. The traditional cascade method suffers from issues such as error propagation. End-to-end models are limited in their ability to handle low-resource languages due to data scarcity. In contrast, large speech models have emerged with their extensive parameter spaces can capture rich features from small datasets and allow fine-tuning for specific tasks. In this study, we focused on the Tibetan-Chinese speech-to-speech translation task and proposed a selective freezing fine-tuning technique for the large speech model to optimize it for specific tasks. At the same time, we propose a LoRA fine-tuning technique based on the selective freezing method to further enhance the model's performance. The results demonstrate that the large speech model-based approach exhibits superior performance, showcasing its robust capability in handling low-resource language translation tasks.

Keywords: Tibetan-Chinese speech-to-speech translation; large speech models; model fine-tuning; end-to-end models; cascade models

1. Introduction

Language is the most important tool for human communication. Speech-to-Speech Translation (S2ST) facilitates communication between speakers of different languages and is widely used in international interactions, conference interpretation, and film dubbing [1]. The development of S2ST has evolved from traditional cascaded models to end-to-end direct speech translation and large speech models [2].

Traditional cascaded speech translation systems consist of Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) synthesis [3, 4, 5]. These models, while benefiting from pre-optimized modules and extensive training data, suffer from error propagation, domain mismatches, and high latency. End-to-end models, such as Translatotron [6], leverage attention-based sequence-to-sequence architectures, enabling direct speech-to-speech translation. Advancements like Vector-Quantized Variational Auto-Encoder (VQ-VAE) [7, 8, 9],

[†]Corresponding author: Yue Zhao (Email: zhaoyueso@muc.edu.cn; ORCID:0000–0002-4007-7016)

Speech-to-Unit Translation (S2UT) [10], and HuBERT-based unit clustering [11, 12, 13] further improve performance. Despite benefits like reduced latency and error propagation, data scarcity remains a critical challenge. Large speech models like Whisper [14], MMS [15], AudioPaLM [16], and SeamlessM4T [17] integrate multimodal capabilities, enhancing multilingual S2ST performance. These models mitigate issues related to separate systems and limited language coverage while enabling on-demand translation.

Recent research has begun to address S2ST for minority languages, such as Tibetan-Chinese translation. Studies by Xu et al. [18] and Liu et al. [19] have explored end-to-end models and multitask learning approaches, respectively. However, the scarcity of parallel corpora and speech datasets for minority languages poses significant challenges.

To address these challenges, this paper explores fine-tuning methods based on SeamlessM4T-v1 for Tibetan-Chinese S2ST. We conduct experiments using datasets of different sizes to evaluate the model's performance.Additionally, we compare the results of large speech models with end-to-end approaches to assess their effectiveness in Tibetan-Chinese translation tasks. Our findings highlight the strong potential of large speech models in handling low-resource language translation.

2. Our Method

This paper is based on the SeamlessM4T-v1 large speech model, which employs the multitask UnitY architecture and uses self-supervised discrete acoustic units to represent target speech. The block diagram of the model structure used in this paper is shown in Figure 1. Building upon this, the paper introduces a selective freezing fine-tuning method for the SeamlessM4T-v1 model. This method allows for precise control over the freezing of specific model components. It dynamically determines whether to freeze the speech-to-speech (S2T) or text-to-unit (T2U) components, depending on the fine-tuning mode. This approach facilitates more targeted adjustments during the training process. Additionally, the paper presents a joint fine-tuning strategy, which incorporates LoRA fine-tuning on top of the selective freezing method, further improving the model's performance.

Previous research often relied on TTS technology to generate target speech for model training, due to a scarcity of parallel S2ST training data [6, 20, 21]. In contrast, our work employs an authentic S2ST dataset and avoids using synthetic audio.

2.1. Base Model

In 2023, Meta released the large speech model SeamlessM4T, a multi-task learning model that integrates multiple speech processing functions. It uses a Transformer-based architecture that enables efficient task transformation capabilities through parameter sharing and task-specific adaptation layers. In particular, SeamlessM4T-v1 contains 2.3B parameters that decompose the S2ST task into two processes, S2U and U2S, which in turn decompose S2U into two processes, S2TT(Speech-to-Text Translation) and T2U. In S2U, a pre-trained X2T model enhances the optimization of tasks across T2TT, S2TT, and ASR. Its block diagram is shown in Figure 2.

In the study of discrete acoustic units, SeamlessM4T-v1 extracts features from the 35th layer of the XLS-R-1B model. These features are then processed using the K-means algorithm, creating a mapping codebook from speech to discrete units. For the unit-to-speech (U2S) model, SeamlessM4T-v1 employs a multilingual vocoder to synthesize speech from the learned discrete units, as depicted in Figure 3. This model utilizes the HiFi-GAN vocoder [22] and incorporates language embeddings to replicate the acoustic characteristics of specific languages.

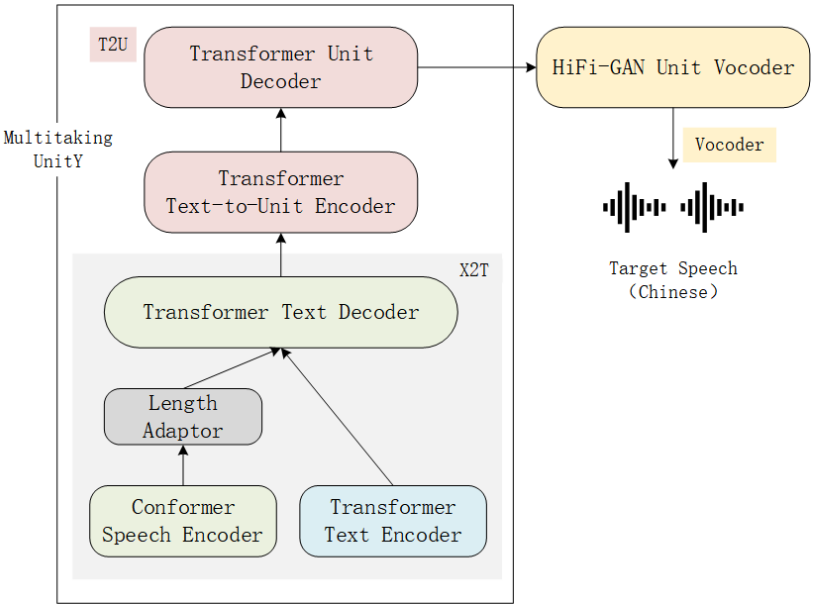

Figure 1: A visual representation of the SeamlessM4T-v1 model architecture and its selective freezing fine-tuning strategy. The figure highlights two key modules eligible for freezing: speech-to-text (S2T, shown in gray) and text-to-unit (T2U, shown in pink). The final output speech is synthesized using the HiFi-GAN vocoder.

We selected SeamlessM4T-v1 as the base model for several reasons: (1) The model is moderate in size, efficient in training, and satisfies our computational resource constraints.

(2) Designed on a multitask learning architecture, the model can handle multiple speech-related tasks simultaneously. This capability enables the model to share learned features when processing speech, improving conversion efficiency and accuracy.

(3) SeamlessM4T-v1 has demonstrated strong performance in other S2ST tasks, showcasing its substantial language understanding capabilities. This performance provides a solid foundation for our downstream tasks.

2.2. Model Fine-tuning

Model fine-tuning is a technique to further train pre-trained models for specific application areas. For the SeamlessM4T-v1 model, this paper conducts an in-depth exploration of its fine-tuning strategies, aiming to achieve more efficient performance optimization for specific tasks. Firstly, we propose an optimization method based on selective freezing. By implementing precise freezing control on certain model components, we dynamically decide whether to freeze the S2T or T2U parts during fine-tuning, as shown in the gray and pink regions of Figure 1. This strategy allows us to flexibly adjust the training focus according to the specific task requirements, thus optimizing the learning process of other modules while maintaining the stability of key modules, effectively improving the model's performance in specific tasks. Given that the S2ST task involves both S2T encoding and T2U generation, we fine-tuned both components in this study to achieve end-to-end performance gains. Building on this, the proposed selective freezing strategy also enables task-specific adaptations for other speech tasks on SeamlessM4T-v1. For example,

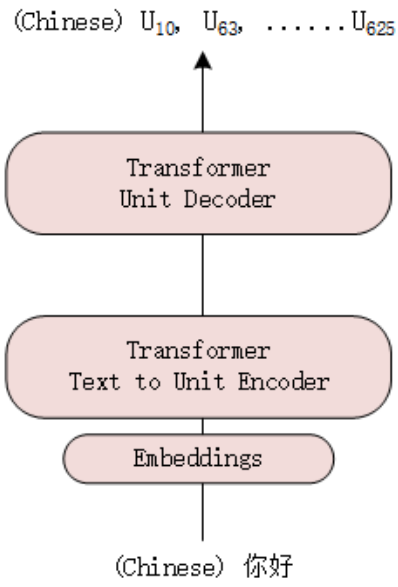

Figure 2: T2U block diagram.

in S2TT fine-tuning, only the S2T module needs to be updated, while the T2U component can remain frozen, thereby reducing computational cost and improving training efficiency.

Secondly, building on the preliminary exploration of the selective freezing method, we further propose a joint fine-tuning strategy. This approach incorporates LoRA fine-tuning technology, optimizing all projection layers involved in the S2ST task. By introducing new trainable low-rank matrices into key parts of the model, LoRA effectively optimizes model parameters. In this study, we apply LoRA modules to all key projection layers involved in the S2ST task. These include the final output projection layer of the text decoder and the output projection layer of the unit decoder. In addition, LoRA is also applied to the query, key, and value matrices within the attention mechanisms of both the Transformer encoder and decoder. This method not only significantly reduces computational overhead during the training process while maintaining the model's expressive capacity, but also enhances the model's generalization ability, enabling it to exhibit stronger adaptability and robustness when facing more complex and diverse tasks.

2.3. Other integrated strategies

In loss function design, we use label smoothing techniques. It transforms hard labels into flexible labels by introducing a smoothing factor. Specifically, the modified label q′(k) is defined as follows:

$$q'(k) = (1 - \epsilon) \cdot \delta(k, y) + \frac{\epsilon}{k} \tag{1}$$

In this equation, ϵ is the smoothing factor and k represents the total number of categories. The function $\delta(k, y)$ is an indicator function that equals 1 when $k = y$ and 0 otherwise.This adjustment reduces the confidence level of the true label category from 1 to $1 - \epsilon$ while distributing ϵ evenly across all non-target categories. Consequently, the label value for each non-target category is elevated from 0 to $\frac{\epsilon}{k}$. This not only helps enhance the model's ability to generalize to new

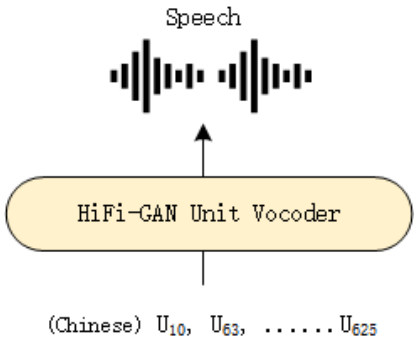

Figure 3: U2S block diagram.

data but also mitigates the problem of overfitting. Label smoothing improves the robustness of the model by adjusting the distribution of the target labels so that the model is not overly confident in any one category during training. Additionally, to ensure stability in a distributed training environment, we aggregated loss data from different computing nodes. This unified loss feedback maintains consistency in model updates.

Moreover, this paper introduces a dynamic learning rate adjustment strategy based on the "Noam" scheduler. This strategy comprises two distinct phases: warm-up and inverse square root decay. During the warm-up phase of model training, the learning rate linearly increases from its initial value to the base learning rate of the optimizer. This process is represented by Equation (2):

$$\mathrm{lr}_t = s + (b - s) \times \frac{t}{T_{\text{warmup}}} \tag{2}$$

In this equation, t represents the current steps, s is the starting learning rate, b denotes the base learning rate, and T_{warmup} stands for the warm-up steps.At the end of the warm-up phase, the learning rate is gradually reduced according to the inverse square root rule, adjusting the formula to Equation (3):

$$\mathrm{lr}_t = b \times \sqrt{\frac{T_{\text{warmup}}}{t}} \tag{3}$$

Here, t also represents the current steps and must be greater than the number of warm-up steps. This learning strategy can quickly achieve model convergence in the early stage of training, and prevent overfitting by fine-tuning the learning rate in the late stage of training. It can also avoid oscillations during the training process to stably optimize the model performance. Additionally, the integration of an early stopping mechanism ceases training prematurely when performance on the validation set is no longer improving. This measure effectively prevents overfitting and saves computational resources.

For the actual training execution, we implemented the Automatic Mixed Precision (AMP) strategy. It significantly improves computational efficiency by using both single and half-precision floating-point numbers during training, while reducing GPU memory consumption. With the support of modern GPU architectures, AMP is able to accelerate the model training process without losing model accuracy. Further, in combination with Distributed Data Parallel (DDP), we parallelized data across multiple GPUs, significantly enhancing the speed and scale of training.

Through the application of these integrated strategies, the fine-tuned SeamlessM4T-v1 model demonstrated superior performance. It provides robust technical support and a solid theoretical foundation for speech translation tasks in low-resource languages.

3. Experiment

3.1. Experimental Data

For the S2ST task based on large speech models, extensive parallel speech data sets are required. To evaluate our proposed methods, we utilized a parallel Tibetan-Chinese speech data set collected by our research team. This data corpus encompasses everyday dialogues and folk culture, including a diverse range of vocabulary such as common words, proper nouns, and geographical names. Furthermore, to validate the robust general knowledge and feature learning capabilities of the large speech models, we employed two different data sets for fine-tuning. Specifically, we selected the Changdu and Yushu languages under the Kham dialect of Tibetan as dataset 1, named it "SmallData." Additionally, we expanded dataset 1 by incorporating the Lhasa language under the Ü-Tsang dialect of Tibetan and the Qinghai languages under the Amdo Tibetan, named it "NormalData." Detailed information on these two datasets is provided in Table 1.

Table 1: Tibetan-Chinese Parallel Speech Data Information

Tibetan-Chinese parallel speech corpus	**SmallData**		**NormalData**	
	Tibetan	**Chinese**	**Tibetan**	**Chinese**
Total duration	2.37h	2.00h	7.32h	7.18h
Number of sentences	1410	1410	3938	3938
Average duration per sentence	4.82	4.22	6.69	6.60
Audio format	.wav	.wav	.wav	.wav
Sampling rate	16kHz	16kHz	16kHz	16kHz

3.2. Experimental Setup

The hyperparameters employed in this experiment are outlined in Table 2.

Table 2: Model Hyperparameter Settings

Parameter Name	**Description**	**Value**
nnodes	Number of nodes	1
nproc-per-node	Number of processes per node	4
learning_rate	Learning rate	$1e^{-6}$
warmup_steps	Number of warm-up steps	100
max_epochs	Maximum training epochs	10
patience	Early stopping patience	5
batch_size	Batch size per training iteration	2
r	LoRA rank (Low-Rank Matrix Dimension)	32
alpha	LoRA scaling factor	64
dropout_p	Probability of dropout (random neuron dropout)	0.2

3.3. Comparison Model

To assess the performance enhancement of the Tibetan-to-Chinese S2ST task based on the large speech model, we selected an end-to-end model proposed by Liu et al. [7] as a comparison model. This model, termed TransMT-S2ST, is a Transformer-based end-to-end multitask learning model specifically designed for Tibetan-to-Chinese S2ST translation. It can be decomposed into three parts, the first part is a S2U model with a speech encoder and a discrete unit decoder, the second part is a speech recognition auxiliary task added during the training process to facilitate the model learning, similar to the Translatotron [6], and the third part is an individually trained vocoder used to convert discrete units into speech waveforms. The block diagram of its model structure is shown in Figure. 4.

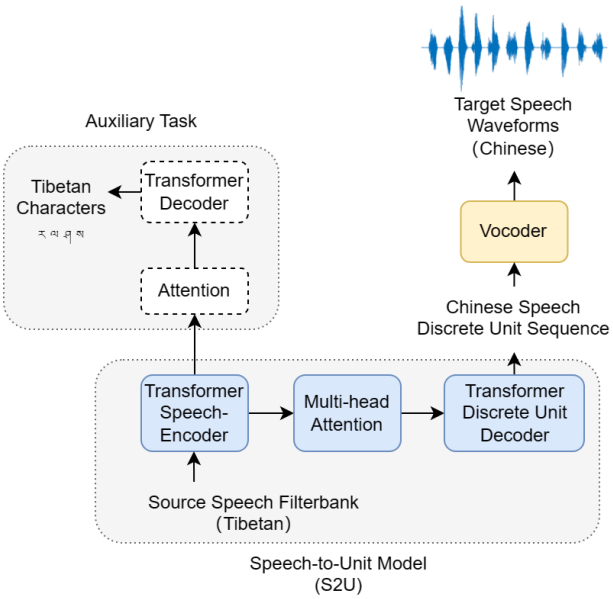

Figure 4: Comparison model architecture [19].

3.4. Experimental Results and Analysis

For the experimental results in this paper, we used two evaluation metrics for evaluation. For objective evaluation metrics, the Bilingual Evaluation Understudy (BLEU) [23] metric was utilized. Table 3 shows the BLEU scores for translating Tibetan speech after fine-tuning SemalessM4T-v1 and the results of translating the same dataset using the comparison model. More specifically, we adopt a tokenization-based BLEU evaluation method to more accurately assess the quality of translation outputs. This approach better captures the semantic and structural similarities between texts. Moreover, it helps avoid the potential biases of character-level BLEU, especially when applied to languages like Chinese, which lack explicit word boundaries. For subjective evaluation, the Mean Opinion Score (MOS) [24] was adopted as the evaluation criterion. We engaged ten native Chinese speakers who were tasked with rating a same set of translated speech results, using the original speech recordings as references. Finally, the average score of all listeners was calculated as the final result. Table 4 shows the results of the subjective evaluation

metrics for this experiment. In Tables 3 and 4, FrSeamlessM4T-v1 denotes the model after the fine-tuning of the freezing method; FrLoSeamlessM4T-v1 denotes the model after the fine-tuning of the freezing combined LoRA method.

Table 3: Objective evaluation results

Dataset	**Model**	**BLEU**		
		Kham Dialect	**Amdo Dialect**	**Ü-Tsang Dialect**
SmallData	TransMT-S2ST	0.02	0.01	0.00
	FrSeamlessM4T-v1	0.47	0.24	0.10
	FrLoSeamlessM4T-v1	**0.52**	**0.26**	**0.13**
NormalData	TransMT-S2ST	0.02	0.01	0.01
	FrSeamlessM4T-v1	0.60	0.64	0.23
	FrLoSeamlessM4T-v1	**0.69**	**0.75**	**0.27**

Table 4: Subjective evaluation results

Dataset	**Model**	**MOS**		
		Kham Dialect	**Amdo Dialect**	**Ü-Tsang Dialect**
SmallData	TransMT-S2ST	3.27	3.15	3.02
	FrSeamlessM4T-v1	3.82	3.77	3.54
	FrLoSeamlessM4T-v1	**3.85**	**3.78**	**3.53**
NormalData	TransMT-S2ST	3.30	3.15	3.10
	FrSeamlessM4T-v1	3.93	3.98	3.81
	FrLoSeamlessM4T-v1	**3.95**	**3.98**	**3.82**

Based on the objective experimental results, the horizontal comparison of BLEU scores reveals that the TransMT-S2ST models trained on both datasets exhibited suboptimal performance. Even with an increased corpus length, there was no significant improvement in model performance. In contrast, the fine-tuned SeamlessM4T-v1 models showed superior results. Specifically, the model fine-tuned on the SmallData dataset performed significantly better in Kham dialect translation. This is primarily due to the fact that all data in the SmallData dataset originates from the Kham dialect, which allows the model to demonstrate better adaptability in translation tasks for this dialect. In comparison, the model fine-tuned on the NormalData dataset showed improvements across all three dialect test sets. This enhancement can be attributed to the increased corpus length, which enables the model to learn richer features and consequently improve the experimental results.

From a vertical analysis, by comparing the performance of SeamlessM4T-v1 models fine-tuned with two different methods and the TransMT-S2ST models, we found that the large models for the Tibetan-to-Chinese S2ST task outperformed the traditional end-to-end models in terms of performance. This result highlights the exceptional capability of large-scale speech models in handling complex speech tasks. Further analysis indicates that even the two SeamlessM4T-v1 models trained with the SmallData dataset performed better across all dialects compared to the TransMT-S2ST model trained with the NormalData dataset. This phenomenon demonstrates the significant advantages of large models in feature learning and general knowledge, showing that even with smaller datasets, large-scale models can still exhibit strong learning and generalization abilities. This also provides a theoretical foundation for the superior performance of large models

on small datasets. In addition, SeamlessM4T-v1 model employs a jointly trained HiFi-GAN neural vocoder, which is optimized end-to-end with the upstream modules. This integration allows for better consistency between the generated speech and the predicted text. In contrast, the TransMT-S2ST model uses a separately trained vocoder, which may introduce a decoupling between text outputs and speech synthesis, potentially affecting the naturalness and clarity of the resulting audio. This difference could also be a contributing factor to the performance gap.

Additionally, experimental results indicate that the performance of the model after combining frozen fine-tuning and LoRA fine-tuning methods surpasses that of using frozen fine-tuning alone. This improvement is due to the LoRA method, which introduces low-rank matrices into key parts of the model, enhancing the model's parameter expression capabilities while reducing computational costs, thus improving training efficiency without compromising model performance.

Despite these improvements, we observe that although the proposed model demonstrates overall performance improvements over the baseline system, the BLEU scores remain relatively low. We attribute this to two main factors. First, the limited duration and scale of the training data may restrict the model's ability to generalize to diverse speech patterns, leading to inaccurate translations for certain utterances. Second, during evaluation, the translated speech must be transcribed into text using an ASR system before BLEU scores can be computed against reference texts. This introduces an additional source of error: even when the translation is semantically correct, variations in prosody, pronunciation, or speaker accent in the generated speech may cause recognition errors.

In terms of subjective evaluation, the fine-tuned SeamlessM4T-v1 model consistently outperformed the TransMT-S2ST model across all dialects. This trend was more pronounced in the NormalData dataset, where the SeamlessM4T-v1 model achieved higher scores, indicating that the translation quality improves with the increase in training data. These results suggest that advanced fine-tuning strategies, such as target adaptation and language feature enhancement implemented in the SeamlessM4T-v1 model, effectively improve translation quality under various dialect and data conditions. However, it is also evident that the MOS score variation between the two fine-tuning methods for the SeamlessM4T-v1 model was minimal. We believe this phenomenon may be related to the use of an integrated neural vocoder. Specifically, our system employs the SeamlessM4T's unified vocoder, which tends to smooth the acoustic output. As a result, perceptual differences in the synthesized speech may be diminished, making it difficult for human listeners to detect subtle variations. This reduced perceptual sensitivity can lead to relatively stable MOS scores, even when the actual translation quality has improved.

4. Conclusion

Speech-to-speech translation technology, as a crucial tool for communication between speakers of different languages, holds significant research value and broad application prospects. This paper explores two fine-tuning strategies—freeze fine-tuning and freeze joint LoRA fine-tuning—based on the recently proposed large speech model SeamlessM4T-v1. We implement freeze control on certain model components while introducing label smoothing techniques and cross-node loss data aggregation to enhance the model's overall performance. Furthermore, through a comparison with traditional end-to-end models, we validate the robustness of large speech models in low-resource language tasks. By analyzing the performance differences of models trained on two different-sized datasets, we further confirm the advantages of large speech

models in acquiring general knowledge and features. In the future, we will delve deeper into different fine-tuning methods to further optimize the model's performance.

Author Contributions

Hetong Fan:
· Methodology: Contributed to integrating large speech models into the Tibetan-Chinese translation task.
· Formal Analysis: Independently conducted error analysis and comparative evaluation of system performance.
· Data Curation: Managed data cleaning, archiving, and version control of speech samples.
· Software: Developed core modules for model training and inference.
· Visualization: Created visual representations to illustrate trends in system performance.
· Writing – Original Draft: Drafted the paper.

Yue Zhao:
· Conceptualization: Proposed the innovative framework of applying large-scale speech models to Tibetan-Chinese speech translation.
· Validation: Designed evaluation procedures and organized cross-validation experiments.
· Resources: Coordinated access to high-performance computing platforms.
· Data Curation: Supervised data annotation standards and ensured corpus consistency.
· Project Administration: Managed project milestones.
· Writing – Review & Editing: Reviewed and refined the manuscript for technical accuracy and logical flow.

Jing Yu:
· Investigation: Conducted in-depth research into the distribution and phonetic characteristics of Tibetan speech data.
· Funding Acquisition: Managed research budget applications.

Zairan Gong:
· Formal Analysis: Performed linguistic-level error categorization and source attribution on translation outputs.
· Validation: Participated in independently reproducing experimental results from multiple optimization cycles.

Xiaona Xu:
· Writing – Review & Editing: Edited the manuscript for academic clarity and coherence.
· Supervision: Provided strategic advice on model deployment and real-world applicability.

Haizhou Li:
· Supervision: Oversaw the research direction and ensured quality of outcomes.

Data Availability

The original data sets are available from the corresponding author on reasonable request.

References

[1] J. Du, M. Zhang, C. Zong, and L. Sun, "Opportunities and challenges for machine translation research in China-summary and prospects for the 8th China workshop on machine translation," Journal of Chinese Information Processing, vol. 27, no. 4, pp. 1–8, 2013.

[2] Z. Yang, "Research on Key Issues of Russian-Chinese Military Speech Translation Based on Seq2Seq Model," Master's thesis, Information Engineering University, 2019.

[3] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan, "JANUS-III: Speech-to-speech translation in multiple languages," in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 99–102, 1997.

[4] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 2, pp. 365–376, 2006.

[5] W. Wahlster, Verbmobil: foundations of speech-to-speech translation, Springer Science and Business Media, 2013.

[6] X. Xu, J. Tan, and Y. Zhao, "End-to-End Tibetan-Chinese Speech Translation Based on Multi-task and Multi-level Pre-training," in 2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), pp. 852–857, 2023.

[7] R. Liu, Y. Zhao, and X. Xu, "Multi-Task Self-Supervised Learning Based Tibetan-Chinese Speech-to-Speech Translation," in 2023 International Conference on Asian Language Processing (IALP), pp. 45–49, 2023.

[8] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," arXiv preprint arXiv:1904.06037, 2019.

[9] A. Tjandra, S. Sakti, and S. Nakamura, "Speech-to-speech translation between untranscribed unknown languages," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 593–600, 2019.

[10] C. Zhang, X. Tan, Y. Ren, T. Qin, K. Zhang, and T.-Y. Liu, "Uwspeech: Speech to speech translation for unwritten languages," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp. 14319–14327, 2021.

[11] A. Van Den Oord, O. Vinyals, and others, "Neural discrete representation learning," Advances in neural information processing systems, vol. 30, 2017.

[12] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, and others, "Direct speech-to-speech translation with discrete units," arXiv preprint arXiv:2107.05604, 2021.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Hubert, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

[14] R. Huang, J. Liu, H. Liu, Y. Ren, L. Zhang, J. He, Z. Zhao, "Transpeech: Speech-to-speech translation with bilateral perturbation," arXiv preprint arXiv:2205.12523, 2022.

[15] K. Wei, L. Zhou, Z. Zhang, L. Chen, S. Liu, L. He, J. Li, and F. Wei, "Joint pre-training with speech and bilingual text for direct speech to speech translation," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in International Conference on Machine Learning, pp. 28492–28518, 2023.

[17] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, and others, "Scaling speech technology to 1,000+ languages," Journal of Machine Learning Research, vol. 25, no. 97, pp. 1–52, 2024.

[18] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. El Badawy, W. Han, E. Kharitonov, and others, "Audiopalm: A large language model that can speak and listen," arXiv preprint arXiv:2306.12925, 2023.

[19] L. Barrault, Y.-A. Chung, M. Cora Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, and others, "SeamlessM4T-Massively Multilingual and Multimodal Machine Translation," arXiv preprint arXiv:2308.11596, 2023.

[20] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," arXiv preprint arXiv:1612.01744, 2016.

[21] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 958–965, 2021.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in neural information processing systems, vol. 33, pp. 17022–17033, 2020.
[23] M. Post, "A call for clarity in reporting BLEU scores," arXiv preprint arXiv:1804.08771, 2018.
[24] International Telecommunication Union. Telecommunication Standardization Sector, Methods for subjective determination of transmission quality, International Telecommunication Union, 1996.